Medical Statistics: Clinical Trials

Nick Fieller
## Contents

### 0. Introduction

0.1 Books ................................................................. 1
0.2 Objectives ......................................................... 2
0.3 Organization of course material .................................. 3
0.4 A Note on R, S-PLUS and MINITAB ............................. 4
0.5 Data sets .............................................................. 5
  0.5.1 R data sets ....................................................... 5
  0.5.2 Data sets in other formats .................................... 6
0.6 R libraries required ................................................ 6
0.6 Outline of Course .................................................. 7

### 1. Background and Basic Concepts

1.1 Definition of Clinical Trial (from Pocock, 1983) .............. 9
1.2 Historical Background .............................................. 10
1.3 Field Trial of Salk Polio Vaccine .................................. 11
1.4 Types of Trial ....................................................... 14
  1.4.1 Further notes: .................................................. 15
1.5 Placebo Effect ....................................................... 18
  1.5.1 Nocebo Effect ................................................... 18
1.6 Blindness of trials .................................................. 18
1.7 Ethical Considerations ............................................. 19
1.8 Publication Ethics .................................................. 21
1.9 Evidence-Based Medicine ......................................... 22
  1.9.1 The Bradford-Hill Criteria .................................... 22
1.10 Summary & Conclusions .......................................... 24
 Tasks 1 .................................................................. 25

### 2. Basic Trial Analysis

2.1 Comments on Tests .................................................. 27
  2.1.1 One-sided and two-sided tests ............................... 27
  2.1.2 Separate and Pooled Variance t-tests ...................... 29
    2.1.2.1 Test equality of variances? .............................. 31
2.2 Parallel Group Designs ............................................. 33
2.3 In series designs .................................................... 34
  2.3.1 Crossover Design ............................................... 36
2.4 Factorial Designs ................................................... 39
2.5 Sequential Designs ................................................ 41
2.6 Summary & Conclusions .......................................... 43
 Tasks 2 .................................................................. 44
3. Randomization........................................................................... 47
  3.1 Simple randomization .......................................................... 47
  3.2 Restricted Randomization .................................................. 50
    3.2.1 Blocking ....................................................................... 50
    3.2.2 Unequal Allocation ....................................................... 52
    3.2.3 Stratified Randomization ............................................. 53
    3.2.4 Minimization .................................................................. 55
      3.2.4.1 Note: Minimization/Adaptive Randomization .......... 58
  3.3 Why Randomize? ................................................................. 58
  3.4 Historical/database controls ............................................... 59
  3.5 Randomization Software .................................................... 60
  3.6 Summary and Conclusions .................................................. 61
  Tasks 3 ..................................................................................... 62
  Exercises 1 .............................................................................. 64

4. Protocol Deviations.................................................................... 66
  4.1 Protocol ............................................................................... 66
  4.2 Protocol deviations .............................................................. 67
  4.3 Summary and Conclusions .................................................. 73

5. Size of the trial........................................................................... 74
  5.1 Introduction .......................................................................... 74
  5.2 Binary Data .......................................................................... 76
  5.3 Quantitative Data ................................................................... 81
  5.4 One-Sample Tests ................................................................... 84
  5.5 Practical problems ............................................................... 85
  5.6 Computer Implementation .................................................... 86
    5.6.1 Implementation in R ..................................................... 86
      5.6.1.1 Example: test of two proportions ......................... 87
      5.6.1.2 Example: t-test of two means ............................. 87
  5.7 Summary and Conclusions .................................................. 88
  Tasks 4 ..................................................................................... 90
  Exercises 2 .............................................................................. 94

6. Multiplicity and interim analysis .............................................. 95
  6.1 Introduction .......................................................................... 95
    6.1.1 Example: Signs of the Zodiac ....................................... 96
  6.2 Multiplicity ............................................................................ 99
    6.2.1 Fundamentals ............................................................... 99
    6.2.2 Bonferroni Corrections ................................................. 101
      Examples: ........................................................................... 101
    6.2.3 Multiple End-points ..................................................... 103
Clinical Trials; Contents

6.2.4 Cautionary Examples ................................................................. 106
6.3 Subgroup analyses ........................................................................ 107
   6.3.1 Fundamentals ........................................................................ 107
   6.3.2 Example: Zodiac (Con\textsuperscript{t}) ........................................ 109
   6.3.3 More Cautionary Examples ....................................................... 112
6.4 Interim analyses ............................................................................ 114
   6.4.1 Fundamentals ........................................................................ 114
   6.4.2 Remedy: ................................................................................ 115
   6.4.3.1 Notes: .............................................................................. 119
   6.4.3.2 Further Notes: .................................................................... 121
6.5 Repeated Measures ........................................................................ 122
   6.5.1 Fundamentals ........................................................................ 122
6.6 Miscellany ....................................................................................... 124
   6.6.1 Regrouping ............................................................................ 124
   6.6.2 Multiple Regression ................................................................ 125
   6.6.2.1 Example: shaving & risk of stroke ...................................... 126
6.7 Summary and Conclusions .............................................................. 127

7. Crossover Trials .................................................................................. 131
   7.1 Introduction ................................................................................ 131
   7.2 Illustration of different types of effects ........................................ 132
   7.3 Model ......................................................................................... 134
   7.3.1 Carryover effect ..................................................................... 136
   7.3.1.1 Notes .............................................................................. 138
   7.3.2 Treatment & period effects ...................................................... 139
   7.3.2.1 Treatment test ................................................................... 139
   7.3.2.2 Period test ......................................................................... 141
7.4* Analysis with Linear Models ........................................................ 142
   7.4.0 Introduction .......................................................................... 142
   7.4.1* Fixed effects analysis ............................................................ 143
   7.4.2* Random effects analysis ........................................................ 145
   7.4.3* Deferment of example ........................................................... 145
7.5 Notes ............................................................................................. 146
7.6 Binary Responses ............................................................................ 149
   7.6.1 Example: (Senn, 2002) ............................................................ 149
7.7 Summary and Conclusions .............................................................. 152
Tasks 5 ............................................................................................... 154
Exercises 3 ......................................................................................... 156
10. Comparing Methods of Measurement ......................... 197
   10.1 Introduction ........................................................................................................... 197
   10.1 Bland & Altman Plots ......................................................................................... 198
   10.2 The Kappa Statistic for Categorical Variables .................................................. 201
       10.3 Examples ......................................................................................................... 202
           10.3.1 Two Categories ....................................................................................... 202
           10.3.2 More than Two Categories ...................................................................... 203
   10.4 Further Modifications ......................................................................................... 204
   10.5 Summary and Conclusions ................................................................................. 205
Notes & Solutions for Tasks 1 .................................................................................... 206
Notes & Solutions for Tasks 2 .................................................................................... 208
Notes & Solutions for Tasks 3 .................................................................................... 214
Notes & Solutions for Tasks 4 .................................................................................... 222
Notes & Solutions for Tasks 5 .................................................................................... 229
Notes & Solutions for Tasks 6 .................................................................................... 236
Notes & Solutions for Exercises 1 .............................................................................. 240
Notes & Solutions for Exercises 2 .............................................................................. 245
Notes & Solutions for Exercises 3 .............................................................................. 247
Notes & Solutions for Exercises 4 .............................................................................. 253
Statistical Methods in Clinical Trials

0. Introduction

0.1 Books


Andersen, B. (1990) *Methodological Errors in Medical Research*. Blackwell


Bland, Martin (2000) *An Introduction to Medical Statistics (3rd Ed)*. OUP.


Matthews, J. N. S. (2006), *An Introduction to Randomized Controlled Clinical Trials. (2nd Ed.)* Chapman & Hall


The two texts which are highlighted cover most of the Clinical Trials section of the Medical Statistics module; the first also has material relevant to the Survival Data Analysis section.

† Indicates a book which goes considerably further than is required for this course (Chapter 5) but is also highly relevant for those taking the second semester course MAS6062 Further Clinical Trials.

★ Indicates a book which contains much material that is relevant to this course but it is primarily a book about Medical Statistics and is strongly recommended to those planning to go for interviews for jobs in the biomedical areas (including the pharmaceutical industry)

0.2 Objectives

The objective of this book is to provide an introduction to some of the statistical methods and statistical issues that arise in medical experiments which involve, in particular, human patients. Such experiments are known collectively as clinical trials.

Many of the statistical techniques used in analyzing data from such experiments are widely used in many other areas (e.g. $\chi^2$-tests in contingency tables, t-tests, analysis of variance). Others which arise particularly in medical data and which are mentioned in this course are McNemar’s test, the Mantel-Haenszel test, logistic regression and the analysis of crossover trials.

As well as techniques of statistical analysis, the course considers some other issues which arise in medical statistics — questions of ethics and of the design of clinical trials.
0.3 Organization of course material

The notes in the main Chapters 1 – 10 are largely covered in the two highlighted books in the list of recommended texts above and are supplemented by various examples and illustrations. A few individual sections are marked by a star, *, which indicates that although they are part of the course they are not central to the main themes of the course. The expository material is supplemented by simple ‘quick problems’ (task sheets) and more substantial exercises. These task sheets are designed for you to test your own understanding of the material. If you are not able to complete the tasks then you should go back to the immediately preceding sections (and re-read the relevant section (and if necessary re-read again & …)). Solutions are provided at the end of the book.
0.4 A Note on R, S-PLUS and MINITAB

The main statistical package for this course is R. It is very similar to the copyright package S-PLUS and the command line commands of S-PLUS are [almost] interchangeable with those of R. Unlike S-PLUS, R has only a very limited menu system which covers some operational aspect but no statistical analyses. A brief guide to getting started in R is available from the course homepage.

R is a freely available programme which can be downloaded over the web from http://cran.r-project.org/ or any of the mirror sites linked from there for installation on your own machine. It is available on University networks. R and S-PLUS are almost identical except that R can only be operated from the command line apart from operational aspects such as loading libraries and opening files. Almost all commands and functions used in one package will work in the other. However, there are some differences between them. In particular, there are some options and parameters available in R functions which are not available in S-PLUS. Both S-PLUS and R have excellent help systems and a quick check with help(function) will pinpoint any differences that are causing difficulties.

A key advantage of R over S-PLUS is the large number of libraries contributed by users to perform many sophisticated analyses.

These are updated very frequently and extend the capabilities substantially. If you are considering using the techniques outside this course (e.g. for some other substantial project) then you would be well advised to use R in preference to S-PLUS. Command-line codes for the more substantial analyses given in the notes for this course have been tested in R. In general, they will work in S-PLUS as well but there could
be some minor difficulties which are easily resolved using the help system.

0.5 Data sets

Data sets used in this course are available in a variety of formats on the associated course web page available here.

0.5.1 R data sets

Those in R are given first and they have extensions .Rdata; to use them it is necessary to copy them to your own hard disk. This is done by using a web browser to navigate to the course web, clicking with the right-hand button and selecting ‘save target as…’ or similar which opens a dialog box for you to specify which folder to save them to. Keeping the default .Rdata extension is recommended and then if you use Windows explorer to locate the file a double click on it will open R with the data set loaded and it will change the working directory to the folder where the file is located. For convenience all the R data sets for Medical Statistics are also given in a WinZip file.

**NOTE:** It is not possible to use a web browser to locate the data set on a web server and then open R by double clicking. The reason is that you only have read access rights to the web page and since R changes the working directory to the folder containing the data set write access is required.
0.5.2 Data sets in other formats
Most of the data sets are available in other formats (Minitab, SPSS etc). It is recommended that the files be downloaded to your own hard disk before loading them into any package but in most cases it is possible to open them in the package in situ by double clicking on them in a web browser. However, this is not possible with R.

0.6 R libraries required
Most of the statistical analyses described in this book use functions within the survival package and the MASS package. It is recommended that each R session should start with

```r
library(MASS)
library(survival)
```

The MASS library is installed with the base system of R but you may need to install the survival package before first usage.
0.6 Outline of Course

1. Background:— historical development of statistics in medical experiments. Basic definitions of placebo effect, blindness and phases of clinical trial.

2. Basic trial analysis:— ‘parallel group’ and ‘in series’ designs, factorial designs & sequential designs.

3. Randomization:— simple and restricted, stratified, objectives of randomization.

4. Protocol deviations:— ‘intention to treat’ and ‘per protocol’ analyses.

5. Size of trial:— sample sizes needed to detect clinically relevant differences with specified power.

6. Multiplicity and interim analyses:— multiple significance testing and subgroup analysis, Bonferroni corrections.

7. Crossover trials:— estimation and testing for treatment, period and carryover effects.

8. Combination of trials:— pooling trials and meta analysis, Simpson’s paradox and the Mantel-Haenszel test


1. Background and Basic Concepts

1.1 Definition of Clinical Trial (from Pocock, 1983)

Any form of planned experiment which involves patients and is designed to elucidate the most appropriate treatment of future patients under a given medical condition

Notes:
(i) Planned experiment (not observational study)
(ii) Inferential Procedure — want to use results on limited sample of patients to find out best treatment in the general population of patients who will require treatment in the future.
1.2 Historical Background
(see e.g. Pocock Ch. 2, Matthews Ch. 1)

1537: Treatment of battle wounds:

Treatment A: Boiling Oil [standard]
Treatment B: Egg yolk + Turpentine + Oil of Roses [new]

New treatment found to be better

1741: Treatment of Scurvy, HMS Edinburgh:

Two patients allocated to each of (1) cider; (2) elixi vitriol;
(3) vinegar; (4) nutmeg, (5) sea water; (6) oranges & lemons

(6) produced “the most sudden and visible good effects.”

Prior to 1950s medicine developed in a haphazard way. Medical literature emphasized individual case studies and treatment was copied:— unscientific & inefficient.

Some advances were made (chiefly in communicable diseases) perhaps because the improvements could not be masked by poor procedure.

Incorporation of statistical techniques is more recent.

e.g. MRC (Medical Research Council in the UK) Streptomycin trial for Tuberculosis (1948) was first to use a randomized control.

MRC cancer trials (with statistician Austin Bradford-Hill) first recognizably modern sequence — laid down the [now] standard procedure.
1.3 Field Trial of Salk Polio Vaccine

In 1954 1.8 million young children in the U.S. were in a trial to assess the effectiveness of Salk vaccine in preventing paralysis/death from polio (which affected 1 in 2000).

Certain areas of the U.S., Canada and Finland were chosen and the vaccine offered to all 2\textsuperscript{nd} grade children. Untreated 1\textsuperscript{st} and 3\textsuperscript{rd} grade children used as the control group, a total of 1 million in all.

Difficulties in this ‘observed control’ approach were anticipated:

(a) only volunteers could be used – these tended to be from wealthier/better educated background (i.e. \textit{volunteer bias})

(b) doctors knew which children had received the vaccine and this could (subconsciously) have influenced their more difficult diagnoses (i.e. a problem of \textit{lack of blindness})

Hence a further 0.8 million took part in a randomised double-blind trial simultaneously. Every child received an injection but half these did not contain vaccine:

\[
\begin{tikzpicture}
  \node (random) {random assignment};
  \node[below of=random] (placebo) {placebo (dummy treatment)};
  \node[above of=random] (vaccine) {vaccine};
  \draw[->] (random) -- (vaccine);
  \draw[->] (random) -- (placebo);
\end{tikzpicture}
\]

and child/parent/evaluating physician did not know which.
### Results of Field Trial of Salk Polio Vaccine

<table>
<thead>
<tr>
<th>Study group</th>
<th>Number in group</th>
<th>Number of cases</th>
<th>Rate per 100 000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed control</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vaccinated 2\textsuperscript{nd} grade</td>
<td>221 998</td>
<td>38</td>
<td>17</td>
</tr>
<tr>
<td>Control 1\textsuperscript{st} and 3\textsuperscript{rd} grade</td>
<td>725 173</td>
<td>330</td>
<td>46</td>
</tr>
<tr>
<td>Unvaccinated 2\textsuperscript{nd} grade</td>
<td>123 605</td>
<td>43</td>
<td>35</td>
</tr>
<tr>
<td><strong>Randomized control</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vaccinated</td>
<td>200 745</td>
<td>33</td>
<td>16</td>
</tr>
<tr>
<td>Control</td>
<td>210 229</td>
<td>115</td>
<td>57</td>
</tr>
<tr>
<td>Not inoculated</td>
<td>338 778</td>
<td>121</td>
<td>36</td>
</tr>
</tbody>
</table>

Results from second part conclusive:

(a) incidence in vaccine group reduced by 50%

(b) paralysis from those getting polio 70% less

(c) no deaths in vaccine group (compared with 4 in placebo group)
Results from first part less so – it was noticed that those 2\textsuperscript{nd} grade children NOT agreeing to vaccination had lower incidence than non-vaccinated controls. It could be that:

(a) those 2\textsuperscript{nd} grade children having vaccine are a self-selected high risk group  
   or  
(b) that there is a complex age effect

Whatever the cause, a valid comparison (treated versus control) was difficult. This provides an example of \textit{volunteer bias}.

Thus, this study was [by accident] a comparison between a randomized controlled double-blind clinical trial and a non-randomized open trial. It revealed the superiority of randomised trials which are now regarded as essential to the definitive comparison and evaluation of medical treatments, just as they had been in other contexts (e.g. agricultural trials) since \textasciitilde1900.
1.4 Types of Trial

Typically a new treatment develops through a research programme (at a pharmaceutical company) who test MANY different manufactured/synthesized compounds. Approximately 1 in 10,000 of those synthesized get to a clinical trial stage (initial pre-clinical screening through chemical analysis, preliminary animal testing etc.). Of these, 1 in 5 reach marketing.

The 4 stages of a [clinical] trial programme after the pre-clinical are:--

Phase I trials: Clinical pharmacology & toxicity concerned with drug safety — not efficacy (i.e. not with whether it is effective). Performed on non-patients or volunteers. Aim to find range of safe and effective doses. investigate metabolism of drugs.

n=10 – 50


n= 50 – 100

Phase III trials: Full-scale evaluation of treatment comparison of drug versus control/standard in (large) trial:

n= 100 – 1000

Phase IV trials: Post-marketing surveillance: long-term studies of side effects, morbidity & mortality.

n= as many as possible
1.4.1 Further notes:

Phase I: First objective is to determine an acceptable single drug dosage, i.e. how much drug can be given without causing serious side effects — such information is often obtained from dosage experiments where a volunteer is given increasing doses of the drug rather than a pre-determined schedule.

Phase II: Small scale and require detailed monitoring of each patient.

Phase III: After a drug has been shown to have some reasonable effect it is necessary to show that it is better than the current standard treatment for the same condition in a large trial involving a substantial number of patients. (‘Standard’: drug already on market, want new drug to be at least equally as good so as to get a share of the market)

Note: Almost all [Phase III] trials now are randomized controlled (comparative) studies:

- group receiving new drug
- group receiving standard drug
To avoid bias (subconscious or otherwise), patients must be assigned at random.

(Bias:— May give very ill people the new drug since there is no chance of standard drug working or perhaps because there is more chance of them showing greater improvement, e.g. blood pressure — those with the highest blood pressure levels can show a greater change than those with moderately high levels).

The comparative effect is important. If we do not have a control group and simply give a new treatment to patients, we cannot say whether any improvement is due to the drug or just to the act of being treated (i.e. the placebo effect). Historical controls (i.e. look for records from past years of people with similar condition when they came for treatment) suffer from similar problems since medical care by doctors and nurses improves generally.
In an early study of the validity of controlled and uncontrolled trials, Foulds (1958) examined reports of psychiatric clinical trials:

- in 52 *uncontrolled* trials, treatment was declared ‘successful’ in 43 cases (83%)

- in 20 *controlled* trials, treatment was ‘successful’ in only 5 cases (25%)

  This is SUSPICIOUS.

Beware also of *publication bias*: only publish ‘results’ that say new drug is better, when other studies disagree. Also concern from conflicts of interest — see §1.8 Publication Ethics
1.5 Placebo Effect

One type of control is a placebo or dummy treatment. This is necessary to counter the *placebo effect* — the psychological benefit of being given any treatment/attention at all (used in a comparative study).

1.5.1 Nocebo Effect

Originally *placebo effect* was taken to refer to both pleasant and harmful effects of a treatment believed to be inert but sometimes this is reserved just for pleasant effects and the term *nocebo effect* used to refer to a harmful effect (placebo and nocebo are the Latin for I will please and I will harm respectively). There are anecdotal reports of nocebo effects being surprisingly extreme such as the case of an attempted suicide with placebo pills during a clinical trial which was only averted by emergency medical intervention, see Reeves *et al*, (2007), General Hospital Psychiatry, 29, 275 – 277.

1.6 Blindness of trials

Using placebos allows the opportunity to make a trial *double blind* — i.e. neither the patient nor the doctor knows which treatment was received. This avoids bias from patient or evaluator attitudes.

*Single blind* — *either* patient *or* evaluator blind

In organizing such a trial there is a coded list which records each patient’s treatment. This is held by a co-ordinator & only broken at analysis (or in emergency).

Clearly, blind trials are *only sometimes possible*; e.g. cannot compare a drug treatment with a surgical treatment.
1.7 Ethical Considerations

Specified in Declaration of Helsinki (1964+amendments) consisting of 32 paragraphs, see http://www.wma.net/e/policy/b3.htm.

Ethical considerations can be different from what the statistician would like.

e.g. some doctors do not like placebos — they see it as preventing a possibly beneficial treatment. (How can you give somebody a treatment that you know will not work?). Paragraph 29 and the 2002 Note of Clarification concerns use of placebo-controlled trials.

There is competition between individual and collective ethics — what may be good for a single individual may not be good for the whole population.

It is agreed that it is unethical to conduct research which is badly planned or executed. We should only put patients in a trial to compare treatment A with treatment B if we are genuinely unsure whether A or B is better.

An important feature is that patients must give their consent to be entered (at least generally) and more than this, they must give informed consent (i.e. they should know what the consequences are of taking the possible treatments).

In the UK, local ethics committees monitor and ‘licence’ all clinical trials — e.g. in each hospital or in each city or regional area.
It is also unethical to perform a trial which has little prospect of reaching any conclusion, e.g. because of insufficient numbers of subjects — see later — or some other aspect of poor design. It may also be unethical to perform a trial which has many more subjects than are needed to reach a conclusion, e.g. in a comparative trial if one treatment proves to be far superior then too many may have received the inferior one.
1.8 Publication Ethics

See BMJ Vol 323, p588, 15/09/01. (http://www.bmj.com/)

Editorial published in all journals that are members of the International Committee of Medical Journal Editors (BMJ, Lancet, New England Journal of Medicine, …).

Concern at articles where declared authors have

- not participated in design of study
- had no access to raw data
- little role in interpretation of data
- not had ultimate control over whether study is published

Instead, the sponsors of the study (e.g. pharmaceutical company) have designed, analysed and interpreted the study (and then decided to publish).

A survey of 3300 academics in 50 universities revealed 20% had had publication delayed by at least 6 months at least once in the past 3 years because of pressure from the sponsors of their study.

Contributors must now sign to declare:

- full responsibility for conduct of study
- had access to data
- controlled decision to publish
1.9 Evidence-Based Medicine

This course is concerned with ‘Evidence-Based Medicine (EBM)’ or more widely ‘Evidence-Based Health Care’. The essence of EBM is that we should consider critically all evidence that a drug is effective or that a particular course of treatment improves some relevant measure of well-being or that some environmental factor causes some condition. Unlike abstract areas of mathematics it is never possible to prove that a drug is effective, it is only possible to assess the strength of the evidence that it is. In this framework statistical methodology has a role but not an exclusive one. A formal test of a hypothesis that a drug has no effect can assess the strength of the evidence against this null hypothesis but it will never be able to prove that it has no effect, nor that it is effective. The statistical test can only add to the overall evidence.

1.9.1 The Bradford-Hill Criteria

To help answer the specific question of causality Austen Bradford-Hill (1965) formulated a set of criteria that could be used to assess whether a particular agent (e.g. a medication or drug or treatment regime or exposure to an environmental factor) caused or influenced a particular outcome (e.g. cure of disease, reduction in pain, medical condition).

These are:–
Temporality (effect follows cause)

Consistency (does it happen in different groups of people – both men and women, different countries)

Coherence (do different types of study result in similar conclusions – controlled trials and observational studies)

Strength of association (the greater the effect compared with those not exposed to the agent the more plausible is the association)

Biological gradient (the stronger the agent the greater the effect – does response follow dose)

Specificity (does agent specifically affect something directly connected with the agent)

Plausibility (is there a possible biological mechanism that could explain the effect)

Freedom from bias or confounding factors (a confounding factor is something related to both the agent and the outcome but is not in itself a cause)

Analogous results found elsewhere (do similar agents have similar results)

These 9 criteria are of course inter-related. Bradford-Hill comments ‘none of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be regarded as a *sine qua non*, that is establishing every one of these does not prove cause and effect nor does failure to establish any of them mean that the hypothesis of cause and effect is completely untrue. However, satisfying most of them does add considerably to the evidence.
1.10 Summary & Conclusions

- Clinical trials involve human patients and are **planned** experiments from which wider **inferences** are to be drawn.

- Randomized controlled trials are the only effective type of clinical trial.

- Clinical Trials can be categorized into 4 phases.

- Double or single blind trials are preferable where possible to reduce bias.

- Placebo effects can be assessed by controls with placebo or dummy treatments where feasible.

- Ethical considerations are part of the statisticians responsibility.
Tasks 1

1. Read the article referred to in §1.8, this can be accessed from the web address given there or from the link given in the course web pages. Use the facility on the BMJ web pages to find related articles both earlier and later.

2. Revision of t-tests and non-parametric tests. The data set HoursSleep which can be accessed from the course website gives the results from a cross-over trial comparing two treatments for insomnia. Group 1 had treatment A in period 1 whilst group 2 had B (and then the other treatment in period 2). Use a t-test to assess the differences between the mean numbers of hours sleep on the two treatments in period 1. Compare the p-values obtained using separate and pooled variance options. Next assess the difference in medians of the two groups using a non-parametric Mann-Whitney test. Compare the p-value obtained from this test with those from the two versions of the t-test.

3. Using your general knowledge compare the following two theories against the Bradford-Hill Criteria:

   (i) Smoking causes lung cancer

   (ii) The MMR (mumps, measles and rubella) vaccine given to young babies causes autism in later childhood.
Clinical Trials; Chapter 1: – Background
2. Basic Trial Analysis

2.1 Comments on Tests

Before considering some basic experimental designs used commonly in the analysis of Clinical Trials there are two comments on statistical tests. The first is on the general question of whether to use a one- or two-sided tests, the other is when considering use of a t-test whether to use the separate or pooled version and what about testing for equality of variance first?

2.1.1 One-sided and two-sided tests

Tests are usually two-sided unless there are very good prior reasons, not observation or data based, for making the test one-sided. If in doubt, then use a two-sided test.

This is particularly contentious amongst some clinicians who say:–

“I know this drug can only possibly lower mean systolic blood pressure so I must use a one-sided test of \( H_0: \mu = \mu_0 \) vs \( H_A: \mu < \mu_0 \) to test whether this drug works.”

The temptation to use a one-sided test is that it is more powerful for a given significance level (i.e. you are more likely to obtain a significant result, i.e. more likely to ‘shew’ your drug works). The reason why you should not is because if the drug actually increased mean systolic blood pressure but you had declared you were using a one-sided test for lower alternatives then the rules of the game would declare that you should ignore this evidence and so fail to detect that the drug is in fact deleterious.
One pragmatic reason for always using two-sided tests is that all good editors of medical journals would almost certainly refuse to publish articles based on use of one-sided tests, (or at the very least question their use and want to be assured that the use of one-sided tests had been declared in the protocol [see §4] in advance (with certified documentary evidence).

A more difficult example is suppose there is suspicion that a supplier is adulterating milk with water. The freezing temperature of watered-down milk is lower than that of whole milk. If you test the suspicions by measuring the freezing temperatures of several samples of the milk, should a one- or two-sided test be used? To answer the very specific question of whether the milk is being adulterated by water you should use a one-sided test but what if in fact the supplier is adding cream?

In passing, it might be noted that the issue of one-sided and two-sided tests only arises in tests relating to one or two parameters in only one dimension. With more than one dimension (or hypotheses relating to more than two parameters) there is no parallel of one-sided alternative hypotheses. This illustrates the rather artificial nature of one-sided tests in general.

Situations where a one-sided test is definitely called for are uncommon but one example is in a case of say two drugs A (the current standard and very expensive) and B (a new generic drug which is much cheaper). Then there might be a proposal that the new cheaper drug should be introduced unless there is evidence that it is very much worse than the standard. In this case the model might have the mean response to the two drugs as $\mu_A = \mu_B$ and if low values are ‘bad’, high values ‘good’ then one might test
H₀: μₐ = μₐ against the one-sided alternative Hₐ: μₐ > μₐ and drug B is introduced if H₀ is not rejected. The reason here is that you want to avoid introducing the new drug if there is even weak evidence that it is worse but if it is indeed preferable then so much the better, you are using as powerful a test as you can (i.e. one-sided rather than the weaker two-sided version). However, this example does raise further issues such as how big a sample should you use and so on. The difficulty here is that you will proceed provided there is absence of evidence saying that you should not do so. A better way of assessing the drug would be to say that you will introduce drug B only if you can shew that it is no more than K units worse than drug A. So you would test H₀: μₐ − K = μₐ against Hₐ: μₐ − K < μₐ and only proceed with the introduction of B if H₀ is rejected in favour of the one-sided alternative (of course you need good medical knowledge to determine a sensible value of K). This leads into the area of non-inferiority trials and bioequivalence studies which are beyond the scope of this course but will be considered in the second semester course MAS6062 Further Clinical Trials.

2.1.2 Separate and Pooled Variance t-tests
This is a quick reminder of some issues relating to two-sample t-tests. The test statistic is the difference in sample means scaled by an estimate of the standard deviation of that difference. There are two plausible ways of estimating the variance of that difference. The first is by estimating the variance of each sample separately and then combining the two separate estimates. The other is to pool all the data from the two samples and estimate a
common variance (allowing for the potential difference in means). The standard deviation used in the test statistic is then the square root of this estimate of variance. To be specific, if we have groups of sizes \( n_1 \) and \( n_2 \), means \( \bar{x}_1 \) & \( \bar{x}_2 \) and sample variances \( s_1^2 \) & \( s_2^2 \) of the two samples then the two versions of a 2-sample t-test are:

(i) separate variance: \( t_r = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \), where the degrees of freedom \( r \) is safely taken as \( \min\{n_1,n_2\} \) though S-PLUS, Minitab and SPSS use a more complicated formula (the Welch approximation) which results in fractional degrees of freedom. This is the default version in R (with function \( \text{t.test()} \)) and Minitab but not in many other packages such as S-PLUS.

(ii) pooled variance: \( t_r = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \)

where \( r = (n_1+n_2-2) \).

This version assumes that the variances of the two samples are equal (though this is difficult to test with small amounts of data). This is the default version in S-PLUS.

We will primarily use the first version because if the underlying populations variances are indeed the same then the separate variance estimate is a good [unbiased] estimate of the common variance and the null distribution of the separate variance estimate test statistic is a t-distribution with only slightly more degrees of freedom than given by the Welch approximation in the statistical packages so resulting in a test that is very slightly conservative and very slightly less powerful. However, if you use the pooled
variance estimate when the underlying population variances are unequal then the resulting test statistic has a null distribution that can be a long way from a t-distribution on \((n_1+n_2-2)\) degrees of freedom and so potentially produce wrong results (neither generally conservative nor liberal, neither generally more nor less powerful, just incorrect). Thus it makes sense to use the separate variance estimate routinely unless there are very good reasons to do otherwise. One such exceptional case is in the calculation of sample sizes [see §5.3] where a pooled variance is used entirely for pragmatic reasons and because many approximations are necessary to obtain any answer at all and this one is not so serious as other assumptions made.

The use of a separate variance based test statistic is only possible since the Welch approximation gives such an accurate estimate of the null distribution of the test statistic and this is only the case in two sample univariate tests. In two-sample multivariate tests or in all multi-sample tests (analysis of variance such as ANOVA and MANOVA) there is no available approximation and a pooled variance estimate has to be used.

### 2.1.2.1 Test equality of variances?

It is natural to consider conducting a preliminary test of equality of variances and then on the basis of the outcome of that decide whether to use a pooled or a separate variance estimate. In fact SPSS automatically gives the results of such a test (Levene’s Test — a common alternative would be Bartlett’s) as well as both versions of the two-sample t-test with two p-values, inviting you to choose. The arguments against using such a preliminary test are
(a) tests of equality of variance are very low powered without large quantities of data — appreciate that a non-significant result does not mean that the variances truly are equal only that the evidence for them being different is weak (b) a technical reason that if the form of the t-test is chosen on the basis of a preliminary test using the same data then allowance needs to be made for the conditioning of the t-test distribution on the preliminary test, i.e. the apparent significance level from the second test (– the t-test) is wrong because it does not allow for the result of the first (– test of equality of variance). You should **definitely not** do both tests and choose the one with the smaller p-value [*data snooping*], which is the temptation from SPSS. In practice the values of the test statistics are usually very close but the p-values differ slightly (because of using a different value for the degrees of freedom in the reference t-distribution). In cases where there is a substantial difference then the ‘separate variance’ version is always the correct one.

Thus the general rule is ‘always use a separate variance test’ noting that in S-PLus the default needs to be changed.
2.2 Parallel Group Designs

Compare k treatments by dividing patients at random into k groups — the $n_i$ patients in group $i$ receive treatment $i$.

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>...</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>...</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>...</td>
<td>•</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>...</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number in group:- $n_1$  $n_2$  $n_3$  ...  $n_k$ : $\Sigma n_i = N$

EACH PATIENT RECEIVES 1 TREATMENT

Often $n_i=n$ with $n \times k = N$ (i.e. groups the same size), but not necessarily, e.g.

- treatment 1 = placebo; $n_1 = 10$
- treatment 2 = drug A; $n_2 = 20$
- treatment 3 = drug B; $n_3 = 20$

with difference between A & B of most interest and ‘hopefully’ differences between drug and placebo will be ‘large’.
Note: Comparisons are ‘between’ patients

Possible analyses:

- 2 groups
- >2 groups

Normal data:
- t-test
- 1-way ANOVA

Non-parametric:
- Mann-Whitney
- Kruskal-Wallis

2.3 In series designs

Here each patient receives all k treatments in the same order

Treatment

\[
1 \rightarrow 2 \rightarrow 3 \rightarrow \ldots \ldots \ldots \rightarrow k
\]

\[
1 \rightarrow X \rightarrow X \rightarrow X \rightarrow \ldots \ldots \rightarrow X
\]

\[
2 \rightarrow X \rightarrow X \rightarrow X \rightarrow \ldots \ldots \rightarrow X
\]

\[
\text{patient} \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot
\]

\[
\text{patient} \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot
\]

\[
\text{patient} \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot
\]

\[
n \rightarrow X \rightarrow X \rightarrow X \rightarrow \ldots \ldots \rightarrow X
\]

Problem: Patients are more likely to enter the trial when their disease is most noticeable, and hence more severe than usual, so there is a realistic chance of a trend towards improvement while on trial regardless of therapy, i.e. the later treatments may appear to be better than the earlier ones.
In most cases, patients differ greatly in their response to any treatment and in their initial disease state. So large numbers are needed in parallel group studies if treatment effects are to be detected.

However there is much less variability between measurements taken on the same patient at different times. Comparisons here are ‘within’ patients.

Advantages:
1. Patients can state ‘preferences’ between treatments
2. Might be able to allocate treatments simultaneously e.g. skin cream on left and right hands

Disadvantages
1. Treatment effect might depend on when it is given
2. Treatment effect may persist into subsequent periods and mask effects of later treatments.
3. Withdrawals cause problems
   (i.e. if a patient leaves before trying all treatments)
4. Not universally applicable,
   e.g. drug treatment compared with surgery
5. Can only use for short term effects

Possible analyses:

- Normal data: paired \( t \)-test (on differences)
- Non-parametric: Wilcoxon signed rank test
- >2 groups: 2-way ANOVA
- Friedman’s test
2.3.1 Crossover Design

Problems with ‘period’ or ‘carryover’ or ‘order’ can be overcome by suitable design; e.g. crossover design. Here patients receive all treatments, but not necessarily in the same order. If patients crossover from one treatment to another there may be problems of feasibility and reliability.

For example, is the disease sufficiently stable and is patient cooperation good enough to ensure that all patients will complete the full course of treatments? A large number of dropouts after the first treatment period makes the crossover design of little value and it might be better to use a between-patient analysis (i.e. parallel group) analysis of the results for period 1 only.
Example 1 (from Pocock, p112)

Effect of the drug oxprenolol on stage-fright in musicians.
N = 24 musicians, double blind in that neither the musician nor the assessor knew the order of treatment.

- Day 1: 12 musicians given oxprenolol, 12 given placebo.
- Day 2: The treatment groups are reversed.

Each musician assessed on each day for nervousness and performance quality.

Can produce the data in the form

<table>
<thead>
<tr>
<th>Patient</th>
<th>Oxp</th>
<th>Plac</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(x_1)</td>
<td>(y_1)</td>
<td>(x_1 - y_1)</td>
</tr>
</tbody>
</table>
| 2       | \(x_2\) | \(y_2\) | \(x_2 - y_2\) | use
| ...     | ...  | ...  | .......... | paired |
| ...     | ...  | ...  | .......... |
| 24      | \(x_{24}\) | \(y_{24}\) | \(x_{24} - y_{24}\) |

More typically design is

washout \(\rightarrow\) treatment \(\rightarrow\) washout \(\rightarrow\) treatment

\[ \text{A} \quad \text{B} \]
\[ \text{B} \quad \text{A} \]

(where ‘washout’ is a period with no treatment at all)
Aside: paired t-test is a one-sample t-test on the differences

\[ t_{n-1} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_d^2 / n}} \]

where \( s_d \) is the standard deviation of the differences, i.e. of the \( n \) values \((x_{1,i} - x_{2,i}), i=1,2,\ldots,n\)

**Example 2:**

Plaque removal of mouthwashes

Treatments

<table>
<thead>
<tr>
<th></th>
<th>A — water</th>
<th>B — brand X</th>
<th>C — brand Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>order of treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
</tbody>
</table>

(and perhaps repeat in blocks of six patients)

Note: If it is not possible for each patient to have each treatment use *balanced incomplete block designs*. 
2.4 Factorial Designs

In some situations, it may be possible to investigate the effect of 2 or more treatments by allowing patients to receive combinations of treatments

![Diagram showing a factorial design with two treatments, A and B, and their combinations.]

Suppose we had 40 patients and allocated 10 at random to each combination, then overall 20 have had A and 20 have had B.

Compare this with a parallel group study to compare A and B (and a placebo), then with about 40 patients available we would have 13 in each group (3x13 ≈ 40).

This factorial design might lead to more efficient comparisons, because of 'larger' numbers.

Obviously not always applicable because of problems with interactions of drugs, but these might themselves be of interest.
Types of interaction

- **lines parallel ⇒ no interaction**
  - Drug A increases response by same amount irrespective of whether patient is also taking B or not

- **quantitative interaction**
  - the effect of A is more marked when patient is also taking B

- **qualitative interaction**
  - A increases response when given alone, but decreases response when in combination with B
2.5 Sequential Designs

In its simplest form, patients are entered into the trial in pairs, one receives A, the other B (allocated at random). Test after results from each pair are known.

e.g. simple preference data (i.e. patient says which of A or B is better)

<table>
<thead>
<tr>
<th>pair</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>preference</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>...</td>
</tr>
</tbody>
</table>

need ‘boundary stopping rules’

e.g.
Advantages

1. Detect large differences quickly

2. Avoids ethical problem of fixed size designs (no patient should receive treatment known to be inferior) — but does complicate the statistical design and analysis

Disadvantages

1. Responses needed quickly (before next pair of patients arrive)

2. Drop-outs cause difficulties

3. Constant surveillance necessary

4. Requires pairing of patients

5. Calculation of boundaries is highly complex. With paired success/failure data (taking A as preferable as a ‘success’) the underlying test is based on a binomial calculation but for individual patients with a quantitative response it is based on a t-test calculation with adjustments made for multiple testing and interim analyses on accumulating data, topics which are discussed further in Chapter 6.
2.6 Summary & Conclusions

♦ ‘Always’ use two-sided tests, not one-sided. One-sided tests are almost cheating.

♦ ‘Always’ use a separate variance t-test.

♦ Never perform a preliminary test of equality of variance.

♦ Parallel group designs — different groups of patients receive different treatments, comparisons are *between* patients

♦ In series designs — all patients receive all treatments in sequence, comparisons are *within* patients

♦ Crossover designs — all patients receive all treatments but different subgroups have them in different orders, comparisons are *within* patients

♦ Factorial designs — some patients receive combinations of treatments simultaneously, difficulties if *interactions*, (quantitative or qualitative), comparisons are *between* patients but more available than in series designs

♦ Sequential designs — suitable for rapidly evaluated outcomes, minimizes numbers of subjects when clear differences between treatments

♦ Efficient design of clinical trials is a crucial ethical element contributed by statistical theory and practice
Tasks 2

1) For each of the proposed trials listed below, select the most appropriate study design, allocating one design to one trial. (One = ‘one and only one’!)

**Trial**

A  Comparison of surgery and 3 months radiotherapy in treating lung cancer.
B  Comparison of new and standard drugs for relief from chronic arthritis
C  Use of diet control and drug therapy for cure of hypertension
D  Comparison of absorption speed of new and standard anaesthetics.

**Design**

a  Crossover
b  Parallel Group
c  Sequential
d  Factorial

2) In a recent radio programme an experiment was proposed to investigate whether common garden snails have a homing instinct and return to their ‘home territory’ if they are moved to some distance away. The proposal is that you should collect a number of snails, mark them with a distinctly coloured nail varnish, and place all of them in your
neighbour’s garden. Your neighbour should do likewise (using a different colour) and place their snails in your garden. You and your neighbour should each observe how many snails returned to their own garden and how many stayed in their neighbour’s. (See http://downloads.bbc.co.uk/radio4/so-you-want-to-be-a-scientist/Snail-Swapping-Experiment-Instructions.pdf for full details)

(a) What flaws does the design of this experiment have?
(b) How could the design of the experiment be improved?

(Note: this question is open-ended and there are many possible acceptable answers to both parts. Discussion is intended)

3) On a recent BBC Radio programme (Front Row, Friday 03/10/08, http://www.bbc.co.uk/radio4/arts/frontrow/) there was an interview with Bettany Hughes, a historian, (http://www.bettanyhughes.co.uk/) who was talking about gold (in relation to an exhibition of a gold statue of Kate Moss in the British Museum). She made the surprising statement

"....ingesting gold can cure some forms of cancer."

I would only regard this as true if there has been a randomized controlled clinical trial where one of the treatments was gold taken by mouth and where the measured outcome was cure of a type of cancer.

The task is to find a record of such a clinical trial or else find a plausible source that might explain this historian's rash statement.
4) What evidence is there that taking fish oil helps schoolchildren concentrate?
3. Randomization

3.1 Simple randomization

For a randomized trial with two treatments A and B the basic concept of tossing a coin (heads=A, tails=B) over and over again is reasonable but clumsy and time consuming. Thus people use tables of random numbers (or generate random numbers in a statistical computer package) instead.

To avoid bias in assigning patients to treatment groups, we need to assign them at random. We need a randomization list so that when a patient (eligible!) arrives they can be assigned to a treatment according to the next number on the list.
Using the following random digits throughout as an example (Neave, table 7.1, row 26, col 1)
3 0 4 5 8 4 9 2 0 7 6 2 3 5 8 4 1 5 3 2 . . . .

Ex 3.1

12 patients, 2 treatments A & B
Assign ‘at random’
e.g. decide 0 to 4 → A
5 to 9 → B
⇒ A A A B B A B A A B B A

Randomization lists can be made as long as necessary & one should make the list before the trial starts and make it long enough to complete the whole trial.
Ex 3.2

With 3 treatments A, B, C
decide 1 to 3 → A
4 to 6 → B
7 to 9 → C
0 → ignore
⇒ A B B C B C A C B A A B

In double blind trials, the randomization list is produced centrally & packs numbered 1 to 12 assembled containing the treatment assigned. Each patient receives the next numbered pack when entering the trial. Neither the doctor nor the patient knows what treatment the pack contains — the randomization code is ‘broken’ only at the end of the trial before the analysis starts. Even then the statistician may not be told which of A, B and C is the placebo and which the active treatment.

Disadvantages:– may lack balance (especially in small trials)
  e.g. in Ex 3.1 7A’s, 5B’s
  in Ex 3.2, 4A’s, 5B’s, 3C’s

Advantage:– each treatment is completely unpredictable, and probability theory guarantees that in the long run the numbers of patients on each treatment will not be substantially different.
3.2 Restricted Randomization

3.2.1 Blocking

Block randomization ensures equal treatment numbers at certain equally spaced points in the sequence of patient assignments. Each random digit specifies what treatment is given to the next block of patients.

In Ex 3.1 (12 patients, 2 treatments A & B)

0 to 4 → AB

⇒ AB AB AB BA BA AB BA

5 to 9 → BA

In Ex 3.2 (3 treatments A, B & C)

1 → ABC

2 → ACB

3 → BAC

4 → BCA

5 → CAB

6 → CBA

7,8,9,0 → ignore

⇒ BAC BCA CAB BCA

Disadvantage:– This blocking is easy to crack/decipher and so it may not preserve the double blinding.

With 2 treatments we could use a block size of 4 to try to preserve blindness
Ex 3.3

1 → AABB
2 → ABAB
3 → ABBA
4 → BBAA
5 → BABA
6 → BAAB
7,8,9,0 → ignore

⇒ ABBA BBAA BABA

Problem:– at the end of each block a clinician who keeps track of previous assignments could predict what the next treatment would be, though in double-blind trials this would not normally be possible. The smaller the choice of block size the greater the risk of randomization becoming predictable.

A trial without ‘stratification’ (i.e. all patients of the same ‘type’ or category) should have a reasonably large block size so as to reduce prediction but not so large that stopping in the middle of a block would cause serious inequality.

In stratified randomization one might use random permuted blocks for patients classified separately into several types (or strata) and in these circumstances the block size needs to be quite small.
3.2.2 Unequal Allocation

In some situations, we may not want complete balanced numbers on each treatment but a fixed ratio.

e.g. A Standard

B New ← need most information on this

decide on a fixed ratio of 1:2 ⇒ need blocking

Reason: more accurate estimates for effects of B; A variation probably known reasonably well already if it is the standard.

Identify all the 3!/2! possible orderings of ABB and assign to digits:

1 to 3 → ABB
4 TO 6 → BAB
7 TO 9 → BBA
0 → ignore

⇒ ABB BAB BAB BBA
3.2.3 Stratified Randomization

(Random permuted blocks within strata)

It is desirable that treatment groups should be as similar as possible in regard of patient characteristics:

relevant patient factors

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>stage of disease</th>
<th>site</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&lt;50,&gt;50)</td>
<td>(M,F)</td>
<td>(1,2,3,4)</td>
<td>(arm,leg)</td>
</tr>
</tbody>
</table>

Group imbalances could occur with respect to these factors: e.g. one treatment group could have more elderly patients or more patients with advanced stages of disease. Treatment effects would then be *confounded* with age or stage (i.e. we could not tell whether a difference between the groups was because of the different treatments or because of the different ages or stages).

Doubt would be cast on whether the randomization had been done correctly and it would affect the credibility of any treatment comparisons.
We can allow for this at the analysis stage through regression (or analysis of covariance) models, however we could avoid it by using a stratified randomization scheme. Here we prepare a separate randomization list for each stratum.

e.g. (looking at age and sex) 8 patients available in each stratum

| <50, M | A B B A | B B A A |
| ≥ 50, M | B A B A | B A A B |
| <50, F | A B A B | B A A B |
| ≥ 50, F | A B A B | A B B A |

so as a new patient enters the trial, the treatment assigned is taken from the next available on the list corresponding to their age and sex.
3.2.4 Minimization

If there are many factors, stratification may not be possible. We might then adjust the randomization *dynamically* to achieve balance, i.e. *minimization* (or *adaptive randomization*). This effectively balances the marginal totals for each level of each factor — however, it loses some randomness. The method is to allocate a new patient with a particular combination of factors to that treatment which ‘balances’ the numbers on each treatment with that combination. See example below.

**Ex 3.5 Minimization** (from Pocock, p.85)

Advanced breast cancer, two treatments A & B, 80 patients already in trial. 4 factors thought to be relevant:–

- ‘performance status’ (ambulatory/non-ambulatory),
- ‘age’ (<50/\(\geq 50\)),
- ‘disease free-time’ (<2/\(\geq 2\) years),
- ‘dominant lesion’ (visceral/osseous/soft tissue).

Suppose that 80 subjects have already been recruited to the study. A new patient enters the trial who is *ambulatory*, <50, has \(\geq 2\) years disease free time and a *visceral* dominant tissue. To decide which treatment to allocate her to, look at the numbers of patients with those factors on each treatment: suppose that of the 80 already in the study, 61 are ambulatory, 30 of whom are on treatment A, 31 on B; of the 19 non-ambulatory 10 are on A and 9 on B. Similarly of the 35 aged under 50 18 are on A and 17 on B, etc. (the complete set of numbers in each category are given in the table below). We now calculate a ‘score’:
### Factors

<table>
<thead>
<tr>
<th>Factors</th>
<th>A</th>
<th>B</th>
<th>next patient</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance status:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambulatory</td>
<td>30</td>
<td>31</td>
<td>⇐</td>
</tr>
<tr>
<td>Non-ambulatory</td>
<td>10</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td><strong>Age:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;50</td>
<td>18</td>
<td>17</td>
<td>⇐</td>
</tr>
<tr>
<td>≥ 50</td>
<td>22</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td><strong>Disease free-time:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;2 years</td>
<td>31</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>≥ 2 years</td>
<td>9</td>
<td>8</td>
<td>⇐</td>
</tr>
<tr>
<td><strong>Dominant lesion:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visceral</td>
<td>19</td>
<td>21</td>
<td>⇐</td>
</tr>
<tr>
<td>Osseous</td>
<td>8</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Soft tissue</td>
<td>13</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

To date,  
A score = 30 + 18 + 9 + 19 = 76  
B score = 31 + 17 + 8 + 21 = 77  
⇒ put patient on A  
(to balance up the scores)

(if scores equal, toss a coin or use simple randomization)
Unlike other methods of treatment assignment, one does not simply prepare a randomization list in advance. Instead one needs to keep a continually and up-to-date record of treatment assignments by patient factors. Computer software is available to help with this (see §3.5).

**Problem:** one possible problem is that treatment assignment is determined solely by the arrangement to date of previous patients and involves no random process except when the treatment scores are equal. This may not be a serious deficiency since investigators are unlikely to keep track of past assignments and hence advance predictions of treatment assignments should not be possible.

Nevertheless, it may be useful to introduce an element of chance into minimization by assigning the treatment of choice (i.e. the one with smallest sum of marginal totals or ‘score’) with probability p where p > ½ (e.g. p= ¾ might be a suitable choice).

Hence, before the trial starts one could prepare 2 randomization lists. The first is a simple randomisation list where A and B occur equally often for use only when the 2 treatments have equal scores, the second is a list in which the treatment with the smallest score occurs with probability ¾ while the other treatment occurs with probability ¼. Using a table of random numbers this is prepared by assigning S (=Smallest) for digits 1 to 6 and L (=Largest) for digits 7 or 8 (ignore 9 and 0).
3.2.4.1 Note: Minimization/Adaptive Randomization

Note that some authors use the term Adaptive Randomization as a synonym for minimization methods but this is best reserved for situations where the outcomes of the treatment are available before the next subject is randomised and the randomization scheme is adapted to incorporate information from the earlier subjects.

3.3 Why Randomize?

1. To safeguard against selection bias
2. To try to avoid accidental bias
3. To provide a basis for statistical tests
3.4 Historical/database controls

Suppose we put all current patients on new treatment and compare results with records of previous patients on standard treatment. This use of historical controls avoids the need to randomize which many doctors find difficult to accept. It might also lessen the need for a placebo.

Major problems:–

- Patient population may change (no formal inclusion/exclusion criteria before trial started for the historical patients)
- Ancillary care may improve with time ⇒ ‘new’ performance exaggerated.

Database controls suffer from similar problems. We cannot say whether any improvement in patients is due to drug or to act of being treated (placebo effect). It may be possible to use a combination of historical controls supplemented with [a relatively small number of] current controls which serve as a check on the validity of the historical ones.
3.5 Randomization Software

A directory of randomisation software is maintained by Martin Bland at:

http://www-users.york.ac.uk/~mb55/guide/randsery.htm

This includes [free] downloadable programmes for simple and blocked randomization, some commercial software including add-ons for standard packages such as STATA, and links to various commercial randomization services which are used to provide full blinding of trials.

This site also includes some useful further notes on randomization with lists of references etc.

R, S-PLUS and MINITAB provide facilities for random digit generation but this is less easy in SPSS.
3.6 Summary and Conclusions

Randomization

- protects against accidental and selection bias
- provides a basis for statistical tests (e.g. use of normal and t-distributions)

Types of randomization include

- simple (but may be unbalanced over treatments)
- blocked (but small blocks may be decoded)
- stratified (but may require small blocks)
- minimization (but lessens randomness)

Historical and database controls may not reflect change in patient population and change in ancillary care as well as inability to allow for placebo effect.
**Tasks 3**

1) Patients are to be allocated randomly to 3 treatments. Construct a randomization list
   i) for a simple, unrestricted random allocation of 24 patients
   ii) for a restricted allocation stratified on the following factors with 4 patients available in each factor combination:
       Sex: M or F  
       Age: <30; 30≤&<50; ≥50.

2) Patients are to be randomly assigned to active and placebo treatments in the ratio 2:1. To ensure ‘balance’ a block size of 6 is to be used. Construct a randomisation list for a total sample size of 24.

3) Patients are to be randomly assigned to active and placebo treatments in the ratio 3:2. To ensure ‘balance’ a block size of 5 is to be used. Construct a randomisation list for a total sample size of 30

4) 
   i) Fifteen individuals who attend a weightwatchers’ clinic are each to be assigned at random to one of the treatments A, B, C to reduce their weights. Describe and implement a randomized scheme to make a balanced allocation of treatments to individuals.
   ii) Different individuals need to lose differing amounts of weight—as shown below (in pounds).

```
<table>
<thead>
<tr>
<th></th>
<th>1. 27</th>
<th>4. 33</th>
<th>7. 27</th>
<th>10. 24</th>
<th>13. 35</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. 35</td>
<td>5. 23</td>
<td>8. 34</td>
<td>11. 30</td>
<td>14. 36</td>
</tr>
<tr>
<td></td>
<td>3. 24</td>
<td>6. 26</td>
<td>9. 30</td>
<td>12. 39</td>
<td>15. 30</td>
</tr>
</tbody>
</table>
```

Describe and implement a design which makes use of this extra information, and explain why this may give a more illuminating comparison of the treatments.
5) A surgeon wishes to compare two possible surgical techniques for curing a specific heart defect, the current standard and a new experimental technique. 24 patients on the waiting list have agreed to take part in the trial; some information about them is given in the table below.

<table>
<thead>
<tr>
<th>Patient</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>Age</td>
<td>64</td>
<td>65</td>
<td>46</td>
<td>70</td>
<td>68</td>
<td>52</td>
<td>54</td>
<td>52</td>
<td>75</td>
<td>55</td>
<td>50</td>
<td>38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Age</td>
<td>59</td>
<td>56</td>
<td>64</td>
<td>64</td>
<td>41</td>
<td>68</td>
<td>48</td>
<td>63</td>
<td>41</td>
<td>62</td>
<td>49</td>
<td>44</td>
</tr>
</tbody>
</table>

Devise a suitable way of allocating patients to the two treatments, and carry out the allocation.
Exercises 1

1) In the comparison of a new drug A with a standard drug B it is required that patients are assigned to drugs A and B in the proportions 3:1 respectively. Illustrate how this may be achieved for a group of 32 patients, and provide an appropriate randomization list. Comment on the rationale for selecting a greater proportion of patients for drug A.

2) The table below gives the age (≤55/>55), gender (M/F), disease stage (I/II/III) of subjects entering a randomized controlled clinical trial at various intervals and who are to be allocated to treatment or placebo in approximately equal proportions immediately on entry.

<table>
<thead>
<tr>
<th>order of entry</th>
<th>Age</th>
<th>Gender</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>≤55</td>
<td>F</td>
<td>III</td>
</tr>
<tr>
<td>2</td>
<td>≤55</td>
<td>M</td>
<td>III</td>
</tr>
<tr>
<td>3</td>
<td>≤55</td>
<td>M</td>
<td>I</td>
</tr>
<tr>
<td>4</td>
<td>≤55</td>
<td>F</td>
<td>I</td>
</tr>
<tr>
<td>5</td>
<td>&gt;55</td>
<td>F</td>
<td>II</td>
</tr>
<tr>
<td>6</td>
<td>≤55</td>
<td>F</td>
<td>III</td>
</tr>
<tr>
<td>7</td>
<td>&gt;55</td>
<td>F</td>
<td>I</td>
</tr>
<tr>
<td>8</td>
<td>&gt;55</td>
<td>M</td>
<td>III</td>
</tr>
<tr>
<td>9</td>
<td>≤55</td>
<td>M</td>
<td>III</td>
</tr>
<tr>
<td>10</td>
<td>&gt;55</td>
<td>F</td>
<td>III</td>
</tr>
<tr>
<td>11</td>
<td>≤55</td>
<td>F</td>
<td>III</td>
</tr>
<tr>
<td>12</td>
<td>≤55</td>
<td>M</td>
<td>I</td>
</tr>
<tr>
<td>13</td>
<td>&gt;55</td>
<td>F</td>
<td>I</td>
</tr>
</tbody>
</table>
i) Use a minimization method designed to achieve an overall balance between the factors to allocate these subjects in the order given to the two treatments and provide the resulting list of allocations.

ii) Cross-tabulate the treatment received with each [separate] factor.

iii) Construct a list to allocate the subjects to treatment completely randomly without taking any account of any prognostic factor and compare the balance between treatment groups achieved on each of the factors.
4. Protocol Deviations

4.1 Protocol

The protocol for any trial is a written document containing all details of trial conduct.

♦ It is needed to gain permission to conduct any trial.
♦ It should contain items on

- purpose
- design & conduct. (See Pocock, table 3.1)

- **Purpose:**
  - motivation
  - aims

- **Design & conduct:**
  - patient selection criteria
  - (inclusion/exclusion)
  - treatment schedule
  - number of patients
  - (and why)
  - assignment of patients:—
  - trial design & randomization
  - evaluation of response:—
  - baseline measure
  - principal response
  - subsidiary criteria
  - ‘informed consent’ form
  - monitoring/record forms
  - techniques for analysis
4.2 Protocol deviations

Things always go wrong. A protocol deviation occurs when a patient departs from the defined experimental procedure (e.g. does not meet the inclusion/exclusion criteria [e.g. too young], takes 2 tablets instead of 1, forgets to take medicine, takes additional other medicine,.....).

All protocol deviations should be noted in the report and in the analysis.

Our aim in the analysis is to minimize bias in the treatment comparison of interest, i.e. to ensure treatment comparisons are not affected by factors other than treatment differences.

All protocol violations and major deviations should be recorded as they occur.
**Ex 4.1** Medical Research Council (1966) study of surgery vs. Radiotherapy for *operable* lung cancer.

In group assigned to receive surgery, certain proportion found to have tumours which *could not* be removed (i.e. they were not *operable* and so should not have been included in the trial — they did not meet the inclusion criteria). In the radiotherapy group, there was no opportunity to detect similar patients (so there may or may not have been patients who did not meet the inclusion criteria).

1: surgery  
2: radiotherapy

<table>
<thead>
<tr>
<th></th>
<th>perhaps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>includes</td>
</tr>
<tr>
<td></td>
<td>some</td>
</tr>
<tr>
<td>in fact</td>
<td>inoperables</td>
</tr>
</tbody>
</table>

The only fair comparison is between the groups as randomized, even though not all in group 1 received treatment.

If the inoperable cases (likely to have a poorer expected outcome) were removed from group 1 before analysis, the remainder in the group would have different (and probably lower risk) characteristics than the whole of 1.
This is called pragmatic or ‘intention to treat’ analysis, i.e. include all eligible patients as originally randomized and assigned to treatments.

eligible: the only exclusions are patients found after randomization to violate inclusion criteria, and where this could in principle have been discovered at the time of randomization — i.e. clear mistakes (e.g. patient too young or too old).

The alternative to ‘intention to treat’ analysis is ‘per protocol analysis’ (or ‘on treatment’ analysis) where patients who deviate from the protocol are excluded from the analysis (e.g. if they do not take enough pills during the course of the trial)

Note that the data presented in §1.3 on the field trial of the Salk polio vaccine for the non-randomized part of the study can be subjected to an intention to treat analysis. It was intended that all 2nd grade children would be vaccinated but some of them (in fact more than 35% of them) refused the vaccine. If the treatment is regarded as offering the vaccination and inoculating those who accept (rather than giving the vaccination itself) then the rate for all 2nd grade children could be compared to that for the observed controls.
Comparison of **per protocol** and **intention to treat**

**Intention to treat** ⇒ initial randomization OK, but patients who deviate may give very odd responses since all patients are analysed.

**Per protocol** ⇒ randomization is compromised (i.e. no longer completely valid). Ask whether withdrawal of patient is related to treatment (e.g. did patient forget to take enough pills because the drug was very strong?). If the numbers of patients are reduced there is a loss of power.
EX 4.2 (Pocock pp182—)

Randomized double-blind trial compared
low dose of new antidepressant with
high .............................................and with
a control treatment.

50 patients entered the trial but 15 had to withdraw because of possible side effects.

Results:

<table>
<thead>
<tr>
<th>clinical assessment</th>
<th>low dose</th>
<th>high dose</th>
<th>control</th>
</tr>
</thead>
<tbody>
<tr>
<td>very effective</td>
<td>2</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>effective</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>ineffective</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>total assessed</td>
<td>9</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td><strong>withdrawn</strong></td>
<td><strong>6</strong></td>
<td><strong>8</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td>total randomized</td>
<td>15</td>
<td>20</td>
<td>15</td>
</tr>
</tbody>
</table>

Note It looks as if withdrawals are not random — some other reason (as different proportions withdrew in each case)
Analyses

Taking response as “% very effective”

A: per protocol (i.e. only those assessed)

<table>
<thead>
<tr>
<th></th>
<th>low</th>
<th>high</th>
<th>control</th>
</tr>
</thead>
<tbody>
<tr>
<td>% very effective</td>
<td>22%</td>
<td>67%</td>
<td>43%</td>
</tr>
</tbody>
</table>

⇒ ‘high’ dose produced the highest proportion of ‘very effective’ assessments.

B: Intention to treat (i.e. including patient withdrawals)

but regarding all withdrawals as ‘ineffective’ i.e. worst case scenario.

<table>
<thead>
<tr>
<th></th>
<th>low</th>
<th>high</th>
<th>control</th>
</tr>
</thead>
<tbody>
<tr>
<td>% very effective</td>
<td>13%</td>
<td>40%</td>
<td>40%</td>
</tr>
</tbody>
</table>

⇒ no difference between ‘high’ and ‘control’

In fact 14/15 on control were rated as ‘effective’ or ‘very effective’ which is a significantly higher proportion than on high dose or low dose , (p<0.01 in each case). Thus the conclusions from the trial are completely reversed once withdrawals are taken into account.
4.3 Summary and Conclusions

Protocols specify all aspects of a clinical trial, including:

- trial purpose, patient selection criteria
- methods of design and analysis, including randomization
- numbers of subjects
- techniques for analysis
- informed consent form

Protocol deviations:

- intention to treat analysis — may lose power of comparison since subjects in treatment groups may not be homogeneous

- per protocol analysis— may lead to bias since randomization is compromised, may also lose power by reducing numbers of subjects
5. Size of the trial

5.1 Introduction

What sample sizes are required to have a good chance of detecting clinically relevant differences if they exist?

Specifications required

[0. main purpose of trial]

1. main outcome measure (e.g. \( \mu_A, \mu_B \) estimated by \( \bar{X}_A, \bar{X}_B \))

2. method of analysis (e.g. two-sample t-test)

3. result given on standard treatment (or pilot results)

4. how small a difference is it important to detect? (\( \delta = \mu_A - \mu_B \))

5. degree of certainty with which we wish to detect it

   (power, 1-\( \beta \))
Note

♦ 'non-significant difference' is not the same as 'no clinically relevant difference' exists.

♦ mistakes can occur:

  Type I: false positive; treatments equivalent but result significant ($\alpha$ represents risk of false positive result)

  Type II: false negative; treatments different but result non-significant ($\beta$ represents risk of false negative result)
5.2 Binary Data

Count numbers of ‘Successes’ & ‘Failures’, and look at the case when there are equal numbers on standard and new treatments:

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>F</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard</td>
<td>x₁</td>
<td>n–x₁</td>
<td>n</td>
</tr>
<tr>
<td>new</td>
<td>x₂</td>
<td>n–x₂</td>
<td>n</td>
</tr>
</tbody>
</table>

Model: $X_1 \sim B(n, \theta_1)$ and $X_2 \sim B(n, \theta_2)$ (binomial distributions), where $X_1$ and $X_2$ are the numbers of success on standard and new treatments.

Hypotheses: $H_0: \theta_1 = \theta_2$ vs. $H_1: \theta_1 \neq \theta_2$

(i.e. a 2-sided test of proportions)

Approximations: Take Normal approximation to binomial:

$X_1 \sim N(n\theta_1, n\theta_1(1–\theta_1))$ and $X_2 \sim N(n\theta_2, n\theta_2(1–\theta_2))$

Requirements: take $\alpha = P[\text{type I error}] = \text{level of test} = 5\%$

and $\beta = P[\text{type II error}] = 1 - \text{power at } \theta_2=10\%$
Suppose standard gives 90% success and it is of clinical interest if the new treatment gives 95% success (or better), i.e.

\[ \theta_1 = 0.9 \]

\[ \theta_2 = 0.95 \] (i.e. a 5% improvement)

1- \( \beta = \gamma \) is the power of the test and we decide we want \( \gamma(0.95)=0.9 \) (so we want to be 90% sure of detecting an improvement of 5%)

We have \( (X_2/n - X_1/n) \sim N(\theta_2-\theta_1, [\theta_2(1-\theta_2)+\theta_1(1-\theta_1)]/n) \)

since \( \text{var}(X_2/n - X_1/n) = \text{var}(X_2/n)+\text{var}(X_1/n) \)

\[ = \theta_2(1-\theta_2)/n + \theta_1(1-\theta_1)/n \]

so the test statistic is:

\[ \frac{\left(\frac{X_2}{n} - \frac{X_1}{n}\right) - 0}{\sqrt{\text{var}\left(\frac{X_2}{n} - \frac{X_1}{n}\right)}} \sim N(0,1) \text{ under } H_0 : \theta_1 = \theta_2 \]

and we will reject \( H_0 \) at the 5% level if

\[ \left|\frac{X_2}{n} - \frac{X_1}{n}\right| > 1.96 \sqrt{\frac{2 \times 0.9 \times 0.1}{n}} \]

(remembering \( \theta_1=\theta_2=0.9 \) under \( H_0 \))
The power function of the test is

\[ P[\text{reject } H_0 \mid \text{alternative parameter } \theta_2] = \gamma(\theta_2) = P\{|X_2/n - X_1/n| > 1.96\sqrt{2\times0.9\times0.1/n}\mid \theta_1=0.9, \theta_2\}\]

and we require \( \gamma(0.95) = 0.9 \)

[Note that for \( \theta_2=0.95 \), \( \text{var}(X_2/n)=0.95(1-0.95)/n \) but \( \text{var}(X_1/n)=0.9(1-0.9)/n \) since \( \theta_1=0.9 \) still]

Now

\[ \gamma(0.95)=1-P\{|X_2/n-X_1/n| \leq 1.96\sqrt{2\times0.9\times0.1/n}\mid \theta_1=0.9, \theta_2=0.95\} \]

\[ =1-\left[ \Phi\left\{ \frac{1.96\sqrt{2\times0.9\times0.1/n - 0.05}}{\sqrt{0.95\times0.05 + 0.9\times1/n}} \right\} - \Phi\left\{ \frac{-1.96\sqrt{2\times0.9\times0.1/n - 0.05}}{\sqrt{0.95\times0.05 + 0.9\times1/n}} \right\} \right] \]

and the last term \( \approx \Phi\left\{ -1.96 - \frac{0.05\sqrt{n}}{\sqrt{0.95\times0.05 + 0.9\times1}} \right\} \rightarrow 0 \)

so we require \( \Phi\left\{ \frac{1.96\sqrt{2\times0.9\times0.1 - 0.05\sqrt{n}}}{\sqrt{0.95\times0.05 + 0.9\times1}} \right\} \approx 0.1 \)

i.e. \( n \approx \frac{(0.95\times0.05 + 0.9\times1)}{0.05^2} \left\{ \Phi^{-1}(0.1) - 1.96\sqrt{\frac{0.9\times1}{0.95\times0.05 + 0.9\times1}} \right\}^2 \)

i.e. need around 580 patients in each ‘arm’ of the trial (1,160 in total) or more if drop out rate known. Could inflate these by 20% to allow for losses.
General formula:
\[ n \approx \frac{\theta_2(1 - \theta_2) + \theta_1(1 - \theta_1)}{(\theta_2 - \theta_1)^2} \left\{ \Phi^{-1}(\beta) + \Phi^{-1}(\alpha/2) \right\}^2 \]

\{N.B. both \( \Phi^{-1}(\beta) \) and \( \Phi^{-1}(\alpha/2) < 0 \}\}

\( \theta_1 \) and \( \theta_2 \) are the hypothetical percentage successes on the two treatments that might be achieved if each were given to a large population of patients. They reflect the realistic expectations of goals which one wishes to aim for when planning the trial and do not relate directly to the eventual results.

\( \alpha \) is the probability of saying that there is a ‘significant difference’ when the treatments are really equally effective

(i.e. \( \alpha \) represents the risk of a **false positive** result)

\( \beta \) is the probability of not detecting a significant difference when there really is a difference of magnitude \( \theta_1 - \theta_2 \) (**false negative**).
Notes:

1. Approximation requires
   \[
   \frac{\sqrt{2\theta_1(1-\theta_1)}}{\sqrt{\theta_2(1-\theta_2) + \theta_1(1-\theta_1)}} \approx 1
   \]
   which here = 1.14, so reasonable, — otherwise need to use more complex methods.

2. Machin & Campbell (Blackwell, 1997) provide tables for various \( \theta_1, \theta_2, \alpha \) and \( \beta \). There are also computer programmes available.

3. If we can really justify a 1-sided test (e.g. from a pilot study) then put \( \Phi^{-1}(\alpha/2) \rightarrow \Phi^{-1}(\alpha) \). 1–sided testing reduces the required sample size.

4. For given \( \alpha \) and \( \beta \), \( n \) depends mainly on \((\theta_2 - \theta_1)^2 \) (& is roughly inversely proportional) which means that for fixed type I and type II errors if one **halves** the difference in response rates requiring detection one needs a **fourfold** increase in trial size.

5. Freiman et al (1978) *New England Journal of Medicine* reviewed 71 binomial trials which reported no statistical significance. They found that 63% of them had power < 70% for detecting a 50% difference in success rates. (??unethical to spend money on such trials?? [Pocock])

6. \( N \) depends very much on the choice of type II error such that an increase in power from 0.5 to 0.95 requires about 3 times the number of patients.

7. In practice, the determination of trial size does not usually take account of patient factors which might influence predicted outcome.
5.3 Quantitative Data

(i) Quantitative response — standard has mean $\mu_1$ and new treatment has mean $\mu_2$.

(ii) Two-sample t-test, but assume $n$ large, so use Normal approximation: $X_1 \sim N(\mu_1, \sigma^2/n)$ and $X_2 \sim N(\mu_2, \sigma^2/n)$
assume equal sample sizes $n$ and equal known variance $\sigma^2$.

The test works well in practice provided the variances are not very different.

(iii) Assume $\mu_1$ known

(iv) Want to detect a ‘new’ mean of size $\mu_2$, (or $\delta = \mu_2 - \mu_1$ the difference in mean response that it is important to detect).

(v) Power at $\mu_2$ is $1-\beta$, i.e. $\gamma(\mu_2) = 1-\beta$, the degree of certainty to detect such a difference exists.

Test statistic under $H_0$: $\mu_1=\mu_2$ is $T = \frac{\bar{X}_2 - \bar{X}_1 - 0}{\sqrt{\frac{2\sigma^2}{n}}} \sim N(0,1)$

2-sided $\alpha$ test rejects $H_0$ if $\left| \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{2\sigma^2}{n}}} \right| > -\Phi^{-1}(\frac{\alpha}{2})$
Power function if new mean = \( \mu_2 \) is

\[
\gamma(\mu_2) = 1 - P \left\{ \frac{X_2 - X_1 - 0}{\sqrt{2\sigma^2/n}} \leq -\Phi^{-1}(\frac{\alpha}{2}) \right\} = 1 - \Phi \left\{ -\Phi^{-1}(\frac{\alpha}{2}) - \frac{\mu_2 - \mu_1}{\sqrt{2\sigma^2/n}} \right\} - \Phi \left\{ +\Phi^{-1}(\frac{\alpha}{2}) - \frac{\mu_2 - \mu_1}{\sqrt{2\sigma^2/n}} \right\}
\]

and we require \( \gamma(\mu_2) = 1 - \beta \), i.e. set

\[
\beta = \Phi \left\{ -\Phi^{-1}(\frac{\alpha}{2}) - (\mu_2 - \mu_1)\sqrt{\frac{n}{2\sigma^2}} \right\} - \Phi \left\{ +\Phi^{-1}(\frac{\alpha}{2}) - (\mu_2 - \mu_1)\sqrt{\frac{n}{2\sigma^2}} \right\}
\]

As before, 2\textsuperscript{nd} term \( \rightarrow 0 \) as \( n \rightarrow \infty \)

so we need \( \Phi^{-1}(\beta) \approx -\Phi^{-1}(\frac{\alpha}{2}) - (\mu_2 - \mu_1)\sqrt{\frac{n}{2\sigma^2}} \)

or

\[
n = \frac{2\sigma^2}{(\mu_2 - \mu_1)^2} \left\{ \Phi^{-1}(\beta) + \Phi^{-1}(\frac{\alpha}{2}) \right\}^2
\]
Notes:–

1) All comments in binomial case apply here also.

2) Need to know the variance $\sigma^2$ which is difficult in practice. Techniques which can help determine a reasonable guess at a value for it are:–

i) may be able to look at similar earlier studies,

ii) may be able to run a small pilot study,

iii) may be able to say what the likely maximum and minimum possible responses under standard treatment could be and so calculate the likely maximum possible range and then get an approximate value for $\sigma$ as one quarter of the range. Here the rationale is the recognition that for Normal data an approximate 95% confidence interval is $\mu \pm 2\sigma$ so the difference between the maximum and minimum is roughly $4\sigma$. 
5.4 One-Sample Tests

The two formula given above apply to two-sample tests for proportions (§5.2) and means (§5.3). It is straightforward to derive similar formula for the corresponding one-sample tests.

In the case of a one sample test, the required sample size to achieve a power of \((1-\beta)\) when using a size \(\alpha\) test of detecting a change from a proportion \(\theta_0\) to \(\theta\) is given by

\[
n = \frac{\left\{\Phi^{-1}(\beta)\sqrt{\theta(1-\theta)} + \Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\theta_0(1-\theta_0)}\right\}^2}{(\theta - \theta_0)^2}
\]

In the case of a one sample test on means, the required sample size to achieve a power of \((1-\beta)\) when using a size \(\alpha\) test of detecting a change from a proportion \(\mu_0\) to \(\mu\) is given by

\[
n = \frac{\sigma^2}{(\mu_0 - \mu)^2} \left\{\Phi^{-1}(\beta) + \Phi^{-1}\left(\frac{\alpha}{2}\right)\right\}^2
\]

The prime use of this formula would be in a paired t-test with \(\mu_0=0\).
5.5 Practical problems

1. If recruitment rate of patients is low, it may take a long time to complete trial. This may be unacceptable and may lead to loss of interest. We could
   a) increase $\delta$
   b) relax $\alpha$ and $\beta$
      (and accept that small differences may be missed)
   c) think of using a multicentre trial (see later)

2. Allow for dropouts, missing data, etc.
   e.g. inflate required numbers by 20% to allow for losses

3. Statistical procedures must be as efficient as possible
   — consider more complex designs.
5.6 Computer Implementation

R, S-PLUS and Minitab provide extensive facilities for power and sample size calculations and these are easily found under the Statistics and Stat menus under Power and Sample Size in the last two packages. SPSS does not currently provide any such facilities (i.e. up to version 16). Note that the formulae given above are approximations and so results may differ from those returned by computer packages, perhaps by as much as 10% in some cases. Further, S-PLUS and Minitab use different approximations and continuity corrections. There are many commercial packages available, perhaps the industry standard is nQuery Advisor which has extensive facilities for more complex problems (analysis of variance, regression etc).

The course web page provides a link to small DOS program, POWER.EXE which has good facilities and this can be downloaded from the page. There are also links to other free sources on the web (and a Google search on power sample size will find millions of references). If you use these free programs you should remember how much you have paid for them.

5.6.1 Implementation in R

In R the functions `power.t.test()`, `power.prop.test` and `power.anova.test()` provide the basic calculations needed for finding any one from the remaining two of power, sample size and CRD (referred to as “delta” in R) from the other two in the commonly used statistical tests of means, proportions and one-way analysis of variance. The HELP system provides full details and extensive examples. `power.t.test()` can handle both two-sample and one-sample tests, the former is the default and
the latter requires type="one.sample" in the call to it. 
power.prop.test() only provides facilities for two-sample 
tests. For one-sample the programme power.exe (available from 
the course web page) is available.

5.6.1.1 Example: test of two proportions

Suppose it is wished to determine the sample size required to 
detect a change in proportions from 0.9 to 0.95 in a two sample 
test using a significance level of 0.05 with a power of 0.9 (or 90%).

> power.prop.test(p1=0.9,p2=0.95,power=0.9,sig.level=0.05)
  Two-sample comparison of proportions power calculation
   n = 581.082
   p1 = 0.9
   p2 = 0.95
   sig.level = 0.05
   power = 0.9
   alternative = two.sided
   NOTE: n is number in *each* group

Thus a total sample size of about 1162 is needed, in close 
agreement with that determined by the approximate formula in 
§5.2.

5.6.1.2 Example: t-test of two means

What clinically relevant difference can be detected with a two 
sample t-test using a significance level of 0.05 with power 0.8 (or 
80%) and a total sample size of 150 when the standard deviation 
is 3.6?

> power.t.test(n=75,sd=3.6,power=0.8,sig.level=0.05)
  Two-sample t test power calculation
   n = 75
   delta = 1.657746
   sd = 3.6
   sig.level = 0.05
   power = 0.8
   alternative = two.sided
   NOTE: n is number in *each* group
5.7 Summary and Conclusions

Sample size calculation is ethically important since

- Samples which are too small may have little chance of producing a conclusion, so exposing patients to risk with no outcome
- Samples which are needlessly too large may expose more subjects than necessary to a treatment later found to be inferior

For sample size calculation we need to know

- outcome measure
- method of analysis (including desired significance levels)
- clinical relevant difference
- power
- results on standard treatment (including likely variability)

For practical implementation we need to know the maximum achievable sample size. This could be limited by

- Recruitment rate and time when analysis of results must be performed
- Total size of target population (number of subjects with the condition which is to be the subject of the clinical trial)
- Available budget

In cases where the maximum sample size is limited it is more useful to calculate a table of clinically relevant differences that can be detected with a range of powers using the available sample size.
Sample size facilities in R in the automatically loaded `stats` package are provided by the three functions `power.t.test()`, `power.prop.test()` and `power.anova.test()`. The first handles one and two sample t-tests for equality of means, the second handles two-sample tests on binomial proportions (but not one-sample tests) and the third simple one-way analysis of variance. The first two will calculate any of sample size, power, clinically relevant difference and significance level given values for the other three. The third will calculate the number of groups, the [common] size of each group, the within groups variance, the between groups variance, power and sample size given values for the other five.

Programme `power.exe` (available from the course web pages) will calculate

- one and two-sample t-tests (including paired t-test)
- one and two-sample tests on binomial proportion
- test on single correlation coefficient
- one sample Mann-Whitney U-test
- Mcnemar’s test
- multiple comparisons using 2-sample t-tests
- cross-over trial comparisons
- log rank test (in survival)

Facilities are available in a variety of freeware and commercial software for many more complex analyses (e.g. regression models) though in many practical cases substantial simplification of the intended analysis is required and so calculations can only be used as a guide.
Tasks 4

The commands in R for calculation of power, sample size etc are 
\texttt{power.t.test()} and \texttt{power.prop.test()}. Note that typing the ↑ recalls the last R command and use of Backspace and the ← key allows you to edit the command and run a new version.

1) A trial for the relief of pain in patients with osteoarthritis of the knee is being planned on the basis of a pilot survey which gave a 25% placebo response rate against a 45% active treatment response rate.
   a) How many patients will be needed to be recruited to a trial which in a two-sided 5% level test will detect a difference of this order of magnitude with 90% power? (Calculate this first ‘by hand’ and then using a computer package and compare the answers).
   b) With equal numbers in placebo and active groups, what active rates would be detected with power in the range 50% to 95% and group sizes 60 to 140? (Calculate for power in steps of 15% and group sizes in steps of 20).

2) Woollard & Cooper (1983) Clinical Trials Journal, 20, 89-97, report a clinical trial comparing Moducren and Propranolol as initial therapies in essential hypertension. These authors propose to compare the change in initial blood pressure under the two drugs.
   a) Given that they can recruit only 100 patients in total to the study, calculate the approximate power of the two-sided 5% level t-test which will detect a difference in mean values of $0.5\sigma$, where $\sigma$ is the common standard deviation.
b) How big a sample would be needed in each group if they required a power of 95%? (Calculate this first 'by hand' and then using a computer package and compare the answers).
The commands in R for calculation of power, sample size etc are `power.t.test()` and `power.prop.test()`. Note that typing the ↑ recalls the last R command and use of Backspace and the ← key allows you to edit the command and run a new version.

3) Look at the solutions to Task Sheet 3 and repeat the analyses given there (if you have not already done so).

4) How many subjects are needed to achieve a power of 80% when the standard deviation is 1.5 to detect a difference in two populations means of 0.8 using a two sample t-test? (Note that R gives the number needed in each group, i.e. total is twice number given)

5) How many subjects are needed to achieve a power of 80% when the standard deviation is 1.5 to detect a difference in one population mean from a specified value of 0.8 using a one sample t–test?

6) Do you have an explanation for why the total numbers in Q2 and Q3 are so different?

7) How many subjects are needed to detect a change of 20% from a standard incidence rate of 50% using a two sample test of proportions with a power of 90%?

8) How many subjects are needed to detect a change from 30% to 10% using a two sample test of proportions with a power of 90%?

9) How many subjects are needed to detect a change from 60% to 80% using a two sample test of proportions with a power of 90%?
10) How many subjects are needed to detect a change from 50% to 30% using a two sample test of proportions with a power of 90%?

11) How many subjects are needed to detect a change from 75% to 55% using a two sample test of proportions with a power of 90%?

12) How many subjects are needed to detect a change from 40% to 60% using a two sample test of proportions with a power of 90%?

13) Questions 5, 6, 7, 8, 9 and 10 all involve changes of 20% and a power of 90%. Why are the answers not all identical?

14) Without doing any calculations (neither by hand nor in R) write down the number of subjects needed to detect a change from 45% to 25% using a two sample test of proportions with a power of 90%
Exercises 2

1) In a clinical trial of the use of a drug in twin pregnancies an obstetrician wishes to show a significant prolongation of pregnancy by use of the drug when compared to placebo. She assesses that the standard deviation of pregnancy length is 1.5 weeks, and considers a clinically significant increase in pregnancy length of 1 week to be appropriate.

i) How many pregnancies should be observed to detect such a difference in a test with a 5% significance level and with 80% power?

ii) It is thought that between 40 and 60 pregnancies will be observed to term during the course of the study. What range of increases in length of pregnancy will the study have a reasonable chance (i.e. between 70% and 90%) of detecting?
6. Multiplicity and interim analysis

6.1 Introduction

This section outlines some of the practical problems that arise when several statistical hypothesis tests are performed on the same set of data. This situation arises in many apparently quite different circumstances when analyzing data from clinical trials but the common danger is that the risk of false positive results can be much higher than intended. The particular danger is when the most statistically significant result is selected from amongst the rest for particular attention, perhaps quite unintentionally.

The most common situations where problems of multiplicity (or multiple testing) arise are encountered are

- multiple endpoints
- subgroup analyses
- interim analyses
- repeated measures

The remedies for these problems include adjusting nominal significance levels to allow for the multiplicity (e.g. Bonferroni adjustments or more complex methods in interim analyses), use of special tests (e.g. Tukey’s test for multiple comparisons or Dunnett’s Test for multiple comparisons with a control) or use of more sophisticated statistical techniques (e.g. Analysis of Variance or Multivariate Analysis).
We begin with a brief example (constructed artificially but not far from reality).

### 6.1.1 Example: Signs of the Zodiac

(Effect of new dietary control regime.)

**Data:** 250 subjects chosen ‘randomly’. Weighed at start of week and again at end of week. Data in kg.

**Results:**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight before</td>
<td>250</td>
<td>58.435</td>
<td>12.628</td>
<td>0.799</td>
</tr>
<tr>
<td>Weight after</td>
<td>250</td>
<td>58.309</td>
<td>12.636</td>
<td>0.799</td>
</tr>
<tr>
<td>Difference</td>
<td>250</td>
<td>0.126</td>
<td>1.081</td>
<td>0.068</td>
</tr>
</tbody>
</table>

So, average weight loss is 0.13kg (≈1/4 pound)

Confidence interval for mean weight loss is (−0.009, 0.260)kg.

Paired t-test for weight loss gives a t-statistic of 1.84, giving a p-value of 0.067 (using a two-sided test). (t=0.126/0.068)

**Not quite significant at the 5% level !**

Can anything be done to ‘squeeze’ a significant result out of this expensive study (we’ve been told we cannot change our mind and use a one-sided test instead!) ????

— luckily, the birth dates are available. Perhaps the success of the diet depends upon the personality and determination of the subject. So, look at subgroups of the data by their sign of the Zodiac:—
Mean weight loss by sign of the Zodiac

<table>
<thead>
<tr>
<th>Zodiac sign</th>
<th>n</th>
<th>mean weight loss</th>
<th>standard error of mean</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aquarius</td>
<td>26</td>
<td>0.313</td>
<td>0.217</td>
<td>1.44</td>
<td>0.161</td>
</tr>
<tr>
<td>Aries</td>
<td>15</td>
<td>0.543</td>
<td>0.205</td>
<td>2.65</td>
<td>0.019</td>
</tr>
<tr>
<td>Cancer</td>
<td>21</td>
<td>0.271</td>
<td>0.249</td>
<td>1.09</td>
<td>0.289</td>
</tr>
<tr>
<td>Capricorn</td>
<td>27</td>
<td>-0.191</td>
<td>0.222</td>
<td>-0.86</td>
<td>0.397</td>
</tr>
<tr>
<td>Gemini</td>
<td>18</td>
<td>0.068</td>
<td>0.266</td>
<td>0.26</td>
<td>0.801</td>
</tr>
<tr>
<td>Leo</td>
<td>22</td>
<td>0.194</td>
<td>0.234</td>
<td>0.83</td>
<td>0.416</td>
</tr>
<tr>
<td>Libra</td>
<td>26</td>
<td>0.108</td>
<td>0.217</td>
<td>0.50</td>
<td>0.623</td>
</tr>
<tr>
<td>Pisces</td>
<td>19</td>
<td>0.362</td>
<td>0.232</td>
<td>1.56</td>
<td>0.136</td>
</tr>
<tr>
<td>Sagittarius</td>
<td>12</td>
<td>0.403</td>
<td>0.294</td>
<td>1.37</td>
<td>0.197</td>
</tr>
<tr>
<td>Scorpio</td>
<td>20</td>
<td>0.030</td>
<td>0.274</td>
<td>0.11</td>
<td>0.248</td>
</tr>
<tr>
<td>Taurus</td>
<td>22</td>
<td>-0.315</td>
<td>0.183</td>
<td>-1.72</td>
<td>0.099</td>
</tr>
<tr>
<td>Virgo</td>
<td>22</td>
<td>0.044</td>
<td>0.238</td>
<td>0.18</td>
<td>0.955</td>
</tr>
</tbody>
</table>

Conclusions: those born under the sign of Aries are particularly suited to this new dietary control. It is well known that Arieans have the strength of character and determination to pursue a strict diet and stick to it. On the other hand, there seems to be some suggestion that those under the sign of Taurus have actually put on weight. Again, not really surprising when one considers the typical characteristics of Taurus.............. (& if we also used a 1-sided p-value......... .)

Comment: This is nonsense! The fault arises in that the most significant result was selected for attention without making any allowance for that selection. The subgroups were considered after the first test had proved inconclusive, not before the experiment had been started so the hypothesis that Arieans are good dieters was only suggested by the data and the fact that it gave an apparently significant result. This is almost certainly a false positive result.
**Note:** The data for weight before and weight after were artificially generated as two samples from a Normal distribution with mean 58.5 and variance 12.5, i.e. there should be no significant difference between the mean weights before and after (as indeed there is not). Birth signs were randomly chosen with equal probability. Two sets of data had to be tried before finding this feature of at least one Zodiac sign providing a false positive.

This example will be returned to later, including ways of analysing the data more honestly.
6.2 Multiplicity

6.2.1 Fundamentals

In clinical trials a large amount of information accumulates quickly and it is tempting to analyse many different responses: i.e. to consider multiple end points or perform many hypothesis tests on different combinations of subgroups of subjects.

Be careful!

All statistical tests run the risk of making mistakes and declaring that a real difference exists when in fact the observed difference is due to natural chance variation. However, this risk is controlled for each individual single test and that is precisely what is meant by the significance level of the test or the p-value. The p-value is the more precise calculation of the risk of a false positive result and is more commonly quoted in current literature. The significance level is usually the broader range that the p-value falls or does not fall in, e.g. ‘not significant at the 5% level’ means that the p-value exceeds 0.05 (& may in fact be much larger than 0.05 or possibly only slightly greater).

However, it is difficult to control the overall risk of declaring at least one false positive somewhere if many separate significance tests are performed. If each test is operated at a separate significance level of 5% then we have a 95% chance of not making a mistake on the first test, a 95%×95% (= 90.25%) of avoiding a mistake on either of the first two and so nearly a 10% risk of one or other (or both) of the first two tests resulting in a false positive.
If we perform 10 (independent) tests at the 5% level, then
\[
\text{Prob [reject } H_0 \text{ in at least one test when } H_0 \text{ is true in all cases]} = 1 - (1 - 0.05)^{10} = 0.4
\]
i.e. a 40% chance of declaring a difference when none exists!!!!

Perhaps a more familiar situation is the calculation of Normal Ranges in clinicochemical tests. A ‘normal person’ has been defined as one who has not been sufficiently investigated. A normal range comprise 95% of the values. If 100 normal persons are evaluated by a clinical test then only 95 of them will be declared normal. If they are then subjected to another independent test then only 90 of them will remain as being considered normal. After another 8 tests there will be only 60 normals left.

Aside: A complementary problem is that of false negatives, i.e. failing to detect a difference when one really exists. Clearly the risk diminishes as more and more tests are performed but at the greatly increased risk of more false positives. (If you buy more Lotto tickets you are more likely to win, but at increasing expense). These problems are more complex and are not considered here, nor are they commonly considered in the medical statistical literature.
6.2.2 Bonferroni Corrections

A simple but very conservative remedy to control the risk of making a false positive is to lower the nominal significance level of the individual tests so that when you calculate the overall final risk after performing k tests it turns out to be closer to your intended level, typically 5%. This is known as a *Bonferroni correction*. The simplest form of the rule is that if you want an overall level of $\alpha$ and you perform k (independent) significance tests then each should be run at a nominal $\alpha/k$ level of significance.

Examples:

(a) 5 separate tests will be performed, so to achieve an overall 5% level of significance a result should only be declared if any test is nominally significant at the $5%/5=1\%$ significance level.

(b) 25 tests are to be performed, an overall level of 1% is intended, so each should be run at a nominal level of $1/25=0.04\%$, i.e. a result should not be claimed unless $p<0.0004$ in any one of them.

(c) 12 tests have been performed and the smallest p-value is 0.019. What is the overall level of significance? The Bonferroni method suggests that it is safe to claim only an overall level of $12\times0.019 = 0.228$. Note that this is the situation in the Signs of the Zodiac example above. This suggests we have no worthwhile evidence of any birth sign being particularly suited to dieting. (We will return later to this example).
Note: Clearly, if a large number of tests is to be performed the Bonferroni correction will demand a totally unrealistically small p-value. This is because the Bonferroni method is very conservative — it over-corrects and in part this is because a simple but only roughly approximate formula has been used.

We can make a more exact calculation which says that to achieve a desired overall level of $\alpha$ when performing $k$ tests you should use a nominal level of $\varepsilon$ where $\alpha = 1 - (1 - \varepsilon)^k$, i.e. only declare a result significant at level $\alpha$ if $p < \varepsilon$, where $\varepsilon$ is given by the formula above. It may not appear very easy to calculate the level from this formula and usually it is not worthwhile since it would not really cure the problem of it being over conservative and usually there are better ways of overcoming the problem of multiplicity, by concentrating on the more important objectives of the trial or using a more sophisticated analysis.

Aside: an approximately solution to the formula above is $\varepsilon = \alpha/k$ which is the derivation of the simple Bonferroni correction. The exact solution is $\varepsilon = 1 - \exp\left\{\frac{1}{k}\log(1 - \alpha)\right\}$. 
6.2.3 Multiple End-points

The most common situation where problems of multiple testing arise is when many different outcome measures are used to assess the result of therapy. It is rare that only a single measure is used (‘once you have got hold of the subject then measure everything in sight’). For example, it is routine to record pulse rate, systolic and diastolic blood pressure, perhaps sitting, standing and supine before and after exercise in hypertensive studies. However, separate significance tests on each separate end-point comparison increases the chance of some false positives.

Remedies:

- Bonferroni correction
- Choose primary outcome measure
- Multivariate analysis

Applying Bonferroni corrections is unduly conservative, i.e. it means that you are less likely to be able to declare a real difference exists even if there is one. The reason for this is that the results from multiple outcome measures are likely to be highly correlated. If the drug is successful as judged by standing systolic blood pressure it is quite likely that the sitting systolic blood pressure would provide similar evidence. If you had not measured the other outcomes and so been forced to use a Bonferroni adjustment in multiplying all your p-values by the number of tests and had instead stayed with just the single measure you might have had an interesting result. This would be particularly frustrating if you had considered 20 highly correlated measures,
each providing a nominal p-value of around 0.01 and Bonferroni told you that you could only claim an overall p-value of 0.2.

The recommended remedy is to concentrate on a primary outcome measure with perhaps a few (two or three) secondary measures which you consider as well (perhaps making an informal Bonferroni correction). Of course it is essential that these are decided in advance of the trial and this is stated in the protocol. The choice can be based on medical expertise or from initial results from a pilot study if the trial is a novel situation. This does not preclude recording all measures that you wish but care must be taken in reporting analyses on these — this is particularly true of clinicochemical laboratory results (and especially when they are recorded as within or without ‘Normal Ranges’, see above). Of course these should be scrutinized and any causes for concern reported.

The ideal statistical remedy is to use a multivariate technique though this may require seeking more specialist or professional statistical assistance. Multivariate techniques will make proper allowance in the analysis for correlated observations (e.g. sitting and standing systolic blood pressure). There are multivariate equivalents of routine univariate statistical analyses such as Student’s t-test (it is Hotelling’s $T^2$-test), Analysis of Variance or ANOVA (it is Multivariate Analysis of Variance or MANOVA, with Wilks’ test or the Lawley-Hotelling test).
The advantage of multivariate analysis is that it will handle all measurements simultaneously and return a single p-value assessing the evidence for departure from the null hypothesis, e.g. that there is a difference between the two treatment groups as revealed by the battery of measures. This advantage is balanced by the potential difficulty of interpreting the nature of the difference detected. It may be that all outcome measures ‘are better’ in one group in which case common sense prevails. Practical experience reveals this is often not so simple and experience is needed in interpretation. This is in part the reason that they are perhaps not so widely used in clinical trials. Further, it is not so easy to define criteria of effectiveness in advance for inclusion in a protocol. Many of these multivariate statistical procedures are now included in widely available statistical packages but advice must be to use them with caution unless experienced help is to hand.
6.2.4 Cautionary Examples

Andersen (1990) reports several examples of ignoring the problems of multiplicity. First, (ref: Br J Clin Pharmacol [Suppl.], 1983, 16: 103) a study of the effect of midazolam on sleep in insomniac patients presented a table of $2 \times 9$ tests of significance on measures of platform balance (seconds off balance) made at various times. The case of measuring the same outcome at successive times is a common one which requires a particular form of multivariate analysis termed repeated measures analysis.

Next, (ref: Basic Clin Med 1981, 15: 445) a report of a new compound to treat rheumatoid arthritis evaluated in a double-blind controlled clinical trial, indomethacin being the control treatment. Andersen reports that there were several criteria for effect (i.e. end-points), repeated at various timepoints and various subdivisions. A total of 850 pairwise comparisons were made (t-tests and Fisher's exact test in $2 \times 2$ contingency tables) and 48 of these gave p-values < 0.05. If there were no difference in the treatment groups and 850 tests were made then one might expect that 5% of these would shew ‘significant’ results. 5% of 850 = 850/20 = 42.5 so finding 48 is not very impressive.

Andersen quotes The Lancet (1984, ii: 1457) in relation to measuring everything that you can think of (or ‘casting your net widely’) as saying “Moreover, submitting a larger number of factors to statistical examination not only improves your chances of a positive result but also enhances your reputation for diligence”.

© NRJF, 1996
6.3 Subgroup analyses

6.3.1 Fundamentals

Problems of multiplicity arise when separate comparisons are made within each of several subgroups of the subjects, for example when the sample of patients is subdivided on baseline factors, e.g. on gender and age for example resulting in four subgroups: (i) M>50; (ii) F>50; (iii) M≤50 & (iv) F≤50. Just as with multiple end-points, the chance of picking up an effect when none exists increases with the number of subdivisions.

Often subgroups are quite naturally considered and there are good a priori reasons for investigating them. If so, then this would of course be recorded in the protocol. If the subgroups are only investigated when an overall analysis gives a non-significant result and so subgroups are dredged to retrieve a significant result (as in the Zodiac example) then extreme care is needed to avoid charges of dishonesty. A safe procedure is only to use [post-hoc] subgroup analyses to suggest future hypotheses for testing in a later study.

Remedy:

- Bonferroni adjustments
- Analysis of Variance
- Follow-up tests for multiple comparisons

Bonferroni adjustments can be used but suffer from the same element of conservatism as in other cases but not so acutely since typically tests on separate subgroups are independent (unlike tests on multiple end-points).
The recommended routine remedy is to perform an Analysis of Variance (ANOVA) to investigate differences between the subgroups and then follow up the result of this (if a significant result is detected) to determine which subgroups are ‘interesting’. A one-way analysis of variance can be thought of as a generalisation to several samples of a two-sample t-test to test for the differences between several subgroups. The test examines the null hypothesis that all subgroups have the same mean against the alternative that at least one of them is different from the rest. The rationale for performing this as a preliminary is that if you think that the effect (e.g. a treatment difference) may only be exhibited in one of several subgroups then it means that one (or more) of the subgroups is different from the rest and so it makes sense to examine the statistical evidence for this. Follow-up tests can then be used to identify which one is of interest. There are many possible follow-up tests which are designed to examine slightly different situations. Examples are Tukey’s multiple range test which examines whether the two most different means are ‘significantly different’, Dunnett’s test which examines whether any particular group mean is ‘significantly different’ from a control group, the Neuman-Keuls test which looks to see which pairs of treatments are different and there are many others which may be found in commonly used statistical packages.
6.3.2 Example: Zodiac (Con’t)

Returning yet gain to the signs of the Zodiac example the appropriate analysis when the subjects are classified by Zodiac sign is to perform a one-way analysis of variance of the weight losses with the Zodiac sign as the classification variable. The analysis presented here is performed in MINITAB but other packages would (should) give identical results:

One-way ANOVA: Weight loss versus Zodiac sign

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zodiac s</td>
<td>11</td>
<td>13.44</td>
<td>1.22</td>
<td>1.05</td>
<td>0.405</td>
</tr>
<tr>
<td>Error</td>
<td>238</td>
<td>277.49</td>
<td>1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>249</td>
<td>290.93</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% CIs For Mean Based on Pooled StDev

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>Pooled StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aquarius</td>
<td>26</td>
<td>0.313</td>
<td>1.106</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Aries</td>
<td>15</td>
<td>0.543</td>
<td>0.794</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Cancer</td>
<td>21</td>
<td>0.271</td>
<td>1.140</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Capricorn</td>
<td>27</td>
<td>-0.191</td>
<td>1.155</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Gemini</td>
<td>18</td>
<td>0.068</td>
<td>1.128</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Leo</td>
<td>22</td>
<td>0.194</td>
<td>1.096</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Libra</td>
<td>26</td>
<td>0.108</td>
<td>1.105</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Pisces</td>
<td>19</td>
<td>0.362</td>
<td>1.010</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Sagittarius</td>
<td>12</td>
<td>0.403</td>
<td>1.018</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Scorpio</td>
<td>20</td>
<td>0.030</td>
<td>1.226</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Taurus</td>
<td>22</td>
<td>-0.315</td>
<td>0.860</td>
<td>(--------*--)</td>
</tr>
<tr>
<td>Virgo</td>
<td>22</td>
<td>0.044</td>
<td>1.117</td>
<td>(--------*--)</td>
</tr>
</tbody>
</table>

Pooled StDev = 1.080

This shows that the overall p-value for testing for a difference between the means of the twelve groups is 0.405 >> 0.05 (i.e. non-significant).

The sketch confidence intervals for the means give an impression that the interval for the mean weight loss for Aries just about excludes zero but this makes no allowance for the fact that this is the most extreme of twelve independent intervals. The box plot on the next page gives little indication that any mean is different from zero:
Here the grey boxes indicate inter-quartile ranges (i.e. the ‘middle half’).

At this stage one would stop since there is no evidence of any difference in mean weight loss between the twelve groups but for illustration if we arbitrarily take the final sign (Virgo) as the ‘control’ and use Dunnett’s test to compare each of the others with this then we obtain

Dunnett’s comparisons with a control

- Family error rate = 0.0500        Individual error rate = 0.00599
- Critical value = 2.77:   Control = level (Virgo) of Zodiac sign:
- Intervals for treatment mean minus control mean

<table>
<thead>
<tr>
<th>Level</th>
<th>Lower</th>
<th>Center</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aquarius</td>
<td>-0.598</td>
<td>0.269</td>
<td>1.136</td>
</tr>
<tr>
<td>Aries</td>
<td>-0.503</td>
<td>0.500</td>
<td>1.502</td>
</tr>
<tr>
<td>Cancer</td>
<td>-0.686</td>
<td>0.227</td>
<td>1.141</td>
</tr>
<tr>
<td>Capricorn</td>
<td>-1.095</td>
<td>-0.235</td>
<td>0.625</td>
</tr>
<tr>
<td>Gemini</td>
<td>-0.927</td>
<td>0.024</td>
<td>0.976</td>
</tr>
<tr>
<td>Leo</td>
<td>-0.753</td>
<td>0.150</td>
<td>1.053</td>
</tr>
<tr>
<td>Libra</td>
<td>-0.803</td>
<td>0.064</td>
<td>0.931</td>
</tr>
<tr>
<td>Pisces</td>
<td>-0.620</td>
<td>0.318</td>
<td>1.256</td>
</tr>
<tr>
<td>Sagittarius</td>
<td>-0.716</td>
<td>0.359</td>
<td>1.433</td>
</tr>
<tr>
<td>Scorpio</td>
<td>-0.939</td>
<td>-0.014</td>
<td>0.911</td>
</tr>
<tr>
<td>Taurus</td>
<td>-1.261</td>
<td>-0.359</td>
<td>0.544</td>
</tr>
</tbody>
</table>

### Boxplots of Weight loss by Zodiac sign

(means are indicated by solid circles)
This gives confidence intervals for the difference of each mean from that of the Virgo group, making proper allowance for the multiplicity and it is seen that all of these comfortably include zero so indicating that there is no evidence of any difference when due allowance is made for the multiple comparisons.

Another useful technique in this situation is to look at the twelve p-values associated with the twelve separate tests. If there were any underlying evidence that some groups were shewing an effect then some of them would be clustered towards the lower end of the scale from 0.0 to 1.0 (the values are given in the table on P5).

### Dotplot of p-values

![Dotplot of p-values](image)

This shews that the values are reasonably evenly spread over the range from 0.0 to 1.0 and in particular that the lowest one is not extreme from the rest.
6.3.3 More Cautionary Examples

First, a report of an actual clinical double-blind study where two treatments were compared and there was an extra unusual element of blinding in that in fact the two treatments were actually identical, see Lee, McNear et al (1980), *Circulation*.

1073 patients with coronary heart disease were randomized into group 1 and group 2, baseline factors were reasonably balanced. The response was survival time and on initial analysis the overall differences between treatment groups non-significant.

Then subgroup analyses were performed: 6 groups were identified on the basis of 2 baseline factors (left ventricular contraction pattern:- normal/abnormal; number diseased vessels 1/2/3). A significant difference in survival times was found in one of the groups (abnormal/3, $\chi^2=5.4$, p<0.023) and could be justified scientifically. Sample sizes were quite large:–

\[ n=397: \ n_1=194, \ n_2=203 \]

In fact, all patients were treated in the SAME way — the ‘treatment’ corresponded to the random allocation into 2 groups. Thus a *false positive effect* had been discovered.

“A survey of racial patterns in pernicious anaemia assessed for age distributions (at presentation) in relation to sex and ethnic group (‘European’ origin, black patients and Latin American patients). The statistical method was Student’s t-test. Blacks (p<0.001) and Latin Americans (p<0.05) were younger than ‘Europeans’. However, the significant age differences were confined to the women; the three male groups did not differ significantly from each other. The black women were significantly younger than all the other groups of patients (p<0.001) except the Latin American women and black men, in whom the age difference did not attain statistical significance. Furthermore, a smaller proportion of the black women were 70 years older, and a larger proportion were 40 years or younger than all the other groups. In fact, the age distribution among the black women may be a bimodal one, with one cluster around a median age of 62 and the other around a median age of 31. The Latin American women were not significantly younger than any other group except the ‘European’ men (p<0.05). Within each racial category, the women tended to be younger than the men, but the differences never reached statistical significance.”

It is clear that somewhere in here is evidence of interesting interactions between age, sex and race and a full three-way analysis of variance would elicit this. The p-values clearly make no allowance for multiple testing and it is not clear how many were actually performed since only (almost) the significant ones were reported.

Happily, this paper is many years old and reviewing of medical literature is now much more rigorous and informed, especially from the statistical viewpoint and especially in the New England Journal of Medicine and the BMJ and similar.
6.4 Interim analyses

6.4.1 Fundamentals

It may be desirable to analyse the data from a trial periodically as it becomes available and again problems of multiple testing arise.

Here the remedies are rather different (and considerably more complex) since not only are the sequence of tests not independent but successive tests are based on accumulating data, i.e. the data from the first period test are pooled into that collected subsequently and re-analyzed with the newly obtained values.

The main objectives of this periodic checking are:

- To check protocol compliance, e.g. compliance rate may be very low. Check that investigators are following the trial protocol and quick inspection of each patient’s results provides an immediate awareness of any deviations from intended procedure. If early results indicate some difficulties in the compliance it may be necessary to make alterations in the protocol.
- To pick up bad side effects so that quick action can be taken and warn investigators to look out for such events in future patients.
Feedback:— helps maintain interest in trial and satisfy curiosity amongst investigators. Basic pre-treatment information such as numbers of patients should be available. Overall data on patient response and follow up for all treatments combined can provide a useful idea of how the trial is proceeding.

Detect large treatment effects quickly so one can stop or modify trial.

The primary reason for monitoring trial data for treatment differences is the ethical concern to avoid any patient in the trial receiving a treatment known to be inferior. In addition, one wishes to be efficient in the sense of avoiding unnecessary continuation once the main treatment differences are reasonably obvious.

However, multiplicity problems exist here too. We have repeated significance tests although not independent — so the overall significance level will be much bigger than the nominal level of \( \alpha \) used in each test.

### 6.4.2 Remedy:

To incorporate such interim analyses we must:

- build them into the protocol (e.g. a group sequential design)
- reduce the nominal significance level of each test, so overall level is required \( \alpha \)

However, if we use the standard Bonferroni adjustment then we obtain very conservative procedures for exactly the same reasons
as detailed in earlier sections. Instead we need refined calculations for the appropriate nominal p-values to use at each step to achieve a desired overall significance level. These calculations are different from those given earlier since there the tests were assumed entirely independent; here they assume that the data used for the first test is included in that for the second, both sets in that for the third etc. (i.e. accumulating data) — the exact calculations are complicated. The full details are given in Pocock (1983) and summarized from there in the tables below:

<table>
<thead>
<tr>
<th>Number of repeated tests at the 5% level</th>
<th>overall significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.08</td>
</tr>
<tr>
<td>3</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>0.13</td>
</tr>
<tr>
<td>5</td>
<td>0.14</td>
</tr>
<tr>
<td>10</td>
<td>0.19</td>
</tr>
<tr>
<td>20</td>
<td>0.25</td>
</tr>
<tr>
<td>50</td>
<td>0.32</td>
</tr>
<tr>
<td>100</td>
<td>0.37</td>
</tr>
<tr>
<td>1000</td>
<td>0.53</td>
</tr>
<tr>
<td>∞</td>
<td>1.0</td>
</tr>
</tbody>
</table>
### Nominal significance levels required for repeated two-sided significance testing for various $\alpha$

<table>
<thead>
<tr>
<th>N</th>
<th>$\alpha=0.05$</th>
<th>$\alpha=0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.029</td>
<td>0.0056</td>
</tr>
<tr>
<td>3</td>
<td>0.022</td>
<td>0.0041</td>
</tr>
<tr>
<td>4</td>
<td>0.018</td>
<td>0.0033</td>
</tr>
<tr>
<td>5</td>
<td>0.016</td>
<td>0.0028</td>
</tr>
<tr>
<td>10</td>
<td>0.0106</td>
<td>0.0018</td>
</tr>
<tr>
<td>15</td>
<td>0.0086</td>
<td>0.0015</td>
</tr>
<tr>
<td>20</td>
<td>0.0075</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

Here $N$ is the maximum number of interim analyses to be performed and this is decided in advance (and included in the protocol of course).
6.4.3 Yet More Cautionary Examples

First an example quoted by Pocock (1983, p150). This is a study to compare drug combinations CP and CVP in non-Hodgkins lymphoma. The measure was occurrence or not of tumour shrinkage. The trial was over 2 years and likely to involve about 120 patients. Five interim analyses planned, roughly after every 25th result. The table below gives numbers of 'successes' and nominal p-values using a $\chi^2$ test at each stage.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>CP</th>
<th>CVP</th>
<th>statistic &amp; p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3/14</td>
<td>5/11</td>
<td>1.63 (p&gt;0.20)</td>
</tr>
<tr>
<td>2</td>
<td>11/27</td>
<td>13/24</td>
<td>0.92 (p&gt;0.30)</td>
</tr>
<tr>
<td>3</td>
<td>18/40</td>
<td>17/36</td>
<td>0.04 (p&gt;0.80)</td>
</tr>
<tr>
<td>4</td>
<td>18/54</td>
<td>24/48</td>
<td>3.25 (0.05&lt;p&lt;0.1)</td>
</tr>
<tr>
<td>5</td>
<td>23/67</td>
<td>31/59</td>
<td>4.25 (0.025&lt;p&lt;0.05)</td>
</tr>
</tbody>
</table>

**Conclusion:** Not significant at end of trial (overall p>0.05) since p>0.016, the required nominal value for 5 repeat tests (see table above).
6.4.3.1 Notes:–

- If there had been NO interim analyses and only the final results available then the conclusion would have been different and CVP declared significantly better at the 5% level.

- In the early stages of any trial the response rates can vary a lot and one needs to avoid any over reaction to such early results on small numbers of patients. For instance, here the first 3 responses occurred on CVP but by the time of the first analysis the situation had settled down and the \( \chi^2 \) test showed no significant difference. By the fourth analysis, the results began to look interesting but still there was insufficient evidence to stop the trial. On the final analysis, when the trial was finished anyway, the \( \chi^2 \) test gave \( p=0.04 \) which is not statistically significant, being greater than the required nominal level of 0.016 for \( N=5 \) analyses.

A totally negative interpretation would not be appropriate from these data alone. One could infer that the superiority of the CVP treatment is interesting but not conclusive.
Next, an example quoted by Andersen (1990), (ref: Br J Surg, (1974), 61: 177). "A randomized trial of Trasylol in the treatment of acute pancreatitis was evaluated statistically when 49 patients had been treated. No statistically significant difference was evident between the two groups, but a trend did emerge in favour of one group. The trial was therefore continued. When altogether 100 cases had been treated, the data were analyzed again. There was now a significant difference ($\chi^2 = 4.675$, d.f. = 1, $p< 0.05$) and the trial was published."

In fact the $p$-value is 0.031 and even if only two interim analyses (including the final one) had been planned this is greater than the necessary 0.029 to claim 5% significance.

Continuing to collect data until a significant result is obtained is clearly dishonest — eventually an apparently significant result will be obtained.
6.4.3.2 Further Notes:–

- One decides in advance what is expected as the maximum number of interim analyses and accordingly makes the nominal significance level smaller. e.g. with at most 10 analyses and overall type I error = 0.05 one uses p<0.0106 as the stopping rule at each analysis for a treatment difference. One should also consider whether an overall type I error $\alpha=0.05$ is sufficiently small when considering a stopping rule. There are 2 situations where $\alpha=0.01$ may be more appropriate:
  
i) if a trial is unique in that its findings are unlikely to be replicated in future research studies
  
  ii) if there is more than one patient outcome used in interim analyses and stopping rule is applied to each outcome. However, one possibility would be to have one principal outcome with a stopping rule having $\alpha=0.05$ and have lesser outcomes with $\alpha=0.01$. It has been suggested that a very stringent stopping criterion, say p<0.001, should be used, on the basis that no matter how often one performs interim analyses the overall type I error will remain reasonably small. It also means that the final analysis, if the trial is not stopped early, can be interpreted using standard significance tests without any serious need to allow for earlier repeated testing.

- See Pocock (1983) for more detail.
6.5 Repeated Measures

6.5.1 Fundamentals

Repeated measures arise when the same feature on a patient is measured at several time points, e.g. blood concentration of some metabolite at baseline and then at intervals of 1, 3, 6, 12 and 24 hours after ingestion of a drug. If, for example, there are two groups of subjects (e.g. two treatment groups) it is tempting to use two-sample t-tests on the measures at each time point in sequence. Of course this is incorrect unless adjustments are made. However, diagrams which shew mean values of the two treatment groups plotted against time and which shew error bars for each mean invite the eye to do exactly that and this must be resisted.

Remedies:

- Bonferroni adjustments
- Multivariate analysis for repeated measures
- Construction of summary measures.

No essentially new comments apply to this situation and indeed some examples discussed earlier include a repeated measure element. Bonferroni adjustments are very conservative since the tests will be highly correlated (as with multiple end-points).

Multivariate analysis of repeated measures can take advantage of the fact that the observations are obtained in a sequence and it may be possible to model the correlation structure.
There are special techniques which do this and specialist or professional advice should be sought. Some so-called ‘repeated measures analyses’ in some statistical packages are in fact quite spurious.

Calculation of summary measures includes calculating quantities such as ‘area under the curve’ (AUC) which may have an interpretation as reflecting bioavailability, another is concentrating on change from baseline. As always, the form of the analysis should be fixed before collection of the data.
6.6 Miscellany

6.6.1 Regrouping

The example below illustrates the dangers of post-hoc recombining subgroups, perhaps a complementary problem to that of post-hoc dividing into subgroups. The example is taken from Pocock (1983) who quotes Hjalmarson et al (1981), The Lancet, ii: 823. The table gives the numbers of deaths or survivals in 90 days after acute myocardial infarction with the subgroup for age-group 65-69 combined first with the older subgroup and then with the younger one. For this subgroup the death rates on placebo and metoprolol were 25/174 (14.4%) and 11/165 (6.7%) respectively.

<table>
<thead>
<tr>
<th></th>
<th>placebo</th>
<th>metoprolol</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>deaths</td>
<td>62/697 (8.9%)</td>
<td>40/698 (5.7%)</td>
<td>p&lt;0.02</td>
</tr>
<tr>
<td>age 40–64</td>
<td>26/453 (5.7%)</td>
<td>21/464 (4.5%)</td>
<td>p&gt;0.2</td>
</tr>
<tr>
<td>age 65–74</td>
<td>36/244 (14.8%)</td>
<td>19/234 (8.1%)</td>
<td>p=0.03</td>
</tr>
<tr>
<td><strong>Metoprolol better for elderly?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age 40–69</td>
<td>51/627 (8.1%)</td>
<td>32/629 (5.1%)</td>
<td>p=0.04</td>
</tr>
<tr>
<td>age 70–74</td>
<td>11/70 (15.7%)</td>
<td>8/69 (11.6%)</td>
<td>p&gt;0.2</td>
</tr>
<tr>
<td><strong>Metoprolol better for younger?</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As well as the dangers of multiple testing, this example illustrates the dangers of post-hoc re-grouping, subgroups should be defined on clinical grounds before the data are collected.

Some subgroup effects could be real of course. However, we should only use subgroup analyses to generate future hypotheses.
6.6.2 Multiple Regression

A further situation where multiplicity problems arise in a well-disguised form and which is often ignored is in large regression analyses involving many explanatory variables. This applies whether the model is ordinary regression with a quantitative response or whether it is a logistic regression for success/failure data or even a Cox proportional hazards regression for survival data.

When analysing the results of estimating such models it is usual to look at estimates of the individual coefficients in relation to their standard errors, declare the result ‘significant’ at the 5% level if the ratio is more than 1.96 (or 2) in magnitude and conclude that the corresponding variable ‘is important’ in affecting the response. It is customary for problems of multiplicity to be ignored on the grounds that although there are several or even many separate (non-independent) t-tests involved, each of the variable is of interest in its own right and that is why it was included in the analysis.

However, there are situations where the regression analysis is more of a fishing expedition and it is more a case of ‘let’s plug everything in and see what comes out’, effectively selecting the most significant result for attention. A trap that is all too easy to fall into arises with interactions. Even with a modest number of variables the number of possible pairwise interactions can be large: including all of them in a model ‘to see if any turn out to be significant’ invites a false positive result which can be seriously misleading.
If this is the case then an honest analysis would have to include this feature and make an appropriate correction, such as a Bonferroni one. Interaction terms should only be included where background knowledge indicates they could naturally arise.

### 6.6.2.1 Example: shaving & risk of stroke

In the Autumn of 2003 it was reported widely in the media that men who did not shave regularly were ‘70% more likely to suffer a stroke and 30% more likely to suffer heart disease, according a study at the University of Bristol’. This is an eye-catching item and so was easily accepted as true.

It is likely that these conclusions were based on a logistic regression model, looking at the probability of suffering a stroke, or on some similar regression model. However, it is of importance to know whether firstly there was any *a priori* medical hypothesis that suggested that diligence in shaving was a feature to be investigated and secondly how many other variables were included in the study. The exact reference for this study is Shaving, Coronary Heart Disease, and Stroke: The Caerphilly Study Ebrahim et al. *Am. J. Epidemiol.* 2003; 157: 234-238, see [http://aje.oxfordjournals.org/cgi/content/full/157/3/234](http://aje.oxfordjournals.org/cgi/content/full/157/3/234) and you are invited to read this article critically.
6.7 Summary and Conclusions

Multiplicity can arise in
- testing several different responses
- subgroup analyses
- interim analyses
- repeated measures
- &c.

The effect of multiplicity is to increase the overall risk of a false positive (i.e. the overall significance level).

Problems of multiplicity can be overcome by
- Bonferroni corrections to nominal significance levels
- Other adjustments to nominal significance levels in special cases, e.g. for accumulating data in interim analyses where adjusting for multiplicity can have counter-intuitive effects.
- more sophisticated analyses, e.g. ANOVA or multivariate methods.

Bonferroni adjustments are typically very conservative because in many situations the tests are highly correlated (especially with multiple end-points and repeated measures).

Conservative means ‘safe’ — i.e. you preserve your scientific reputation by avoiding making mistake but at the expense of failing to discover something scientifically interesting.

A final comment is to remember that
“If you torture the data often enough it will eventually confess”
7. Crossover Trials

7.1 Introduction

Where it is possible for patients to receive both treatments under comparison, crossover trials may well be more efficient (i.e. need fewer patients) than a parallel group study.

Recall idea from section 2.: by acting as his/her own control, the effect of large differences between patients can be lessened by looking at within patient comparisons.

Example 7.1 (Pocock, p112)

Hypertension trial:

<table>
<thead>
<tr>
<th></th>
<th>period 1</th>
<th>period 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>washout</td>
<td>→ randomized</td>
<td>(4 weeks)</td>
</tr>
<tr>
<td>randomized</td>
<td></td>
<td>(4 weeks)</td>
</tr>
<tr>
<td>for 4 weeks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\frac{1}{2}$ new drug B</td>
<td>$\frac{1}{2}$ standard A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>systolic blood pressure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>at end of 5 minute exercise test.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B → A</strong>:</td>
<td>55 patients</td>
<td></td>
</tr>
<tr>
<td><strong>A → B</strong>:</td>
<td>54 patients</td>
<td></td>
</tr>
</tbody>
</table>

Possible effect:
- treatment effects $\tau$
- period effect $\pi$
- carryover effect $\lambda$
7.2 Illustration of different types of effects
Note: assuming that 'low' is good throughout

a) Carryover effect

(i) possible explanation:
beneficial effect of B carries over into period 2

Carryover effect
(ii)

Direction of treatment effect different for different periods caused by carryover.

(ii) is more serious, (i) is unlikely to be detected because of low power.
b) **Period effect**

response in period 2 reduced for both treatments, i.e. patients generally improve so period 2 values on average reduced.

c) **treatment effect**

B better than A
7.3 Model

<table>
<thead>
<tr>
<th></th>
<th>period 1</th>
<th>period 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>group 1</td>
<td>A (Y_{11k})</td>
<td>B (Y_{12k})</td>
</tr>
<tr>
<td>group 2</td>
<td>B (Y_{21k})</td>
<td>A (Y_{22k})</td>
</tr>
</tbody>
</table>

response \(Y_{ijk}\) for

- group i (order); \(i=1,2\)
- period j; \(j=1,2\)
- patient k; \(k=1,2,...,n_i\) \((n_1=n_2\) in balanced case)

Effects

- \(\mu\) — overall mean
- \(\tau_A, \tau_B\) — treatment effects
- \(\pi_1, \pi_2\) — period effects
- \(\lambda_A, \lambda_B\) carryover effects (treatment x period interaction)
- \(\alpha_k\) — random patient effect \(\sim N(0,\phi^2)\) (between patients)
- \(\epsilon_{ijk}\) — random errors \(\sim N(0,\sigma^2)\) (independently)

Identifiability

\[
\tau_A + \tau_B = 0 \\
\pi_1 + \pi_2 = 0
\]
If we take expected values, $\alpha_k$ and $\varepsilon_{ijk}$ disappear.

$$Y_{ijk} = \mu + \alpha_k + \pi + \lambda + \varepsilon_{ijk}$$

$$E(Y_{11k}) = \mu + \tau_A + \pi_1$$

$$E(Y_{12k}) = \mu + \tau_B + \pi_2 + \lambda_A$$

To isolate $\tau$, $\pi$ and $\lambda$ effects we consider sums and differences of the $Y_{ijk}$'s.
7.3.1. Carryover effect

Compute $T_{ik} = \frac{1}{2}(Y_{i1k} + Y_{i2k})$ i.e. the average of the 2 values for patient $k$.

Then $T_{1k} \sim N(\mu + \frac{1}{2}\lambda_A, \frac{\sigma^2}{2} + \frac{1}{2}\sigma^2)$ and $T_{2k} \sim N(\mu + \frac{1}{2}\lambda_B, \frac{\sigma^2}{2} + \frac{1}{2}\sigma^2)$

If $\lambda_A = \lambda_B$ i.e. no (differential) carryover, $T_{1k}$ and $T_{2k}$ have identical Normal distributions.

Thus we can test for equality of means of group 1 and group 2 using a 2-sample $t$-test to establish whether

$H_0: \lambda_A = 0 = \lambda_B$ is plausible.

$$\text{i.e. use } \frac{\overline{T}_1 - \overline{T}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_r$$

where $s_i^2$ is the sample variance of the $T_{1k}$ so $\text{var(}\overline{T}_i) = \frac{s_i^2}{n_i}$, etc. and we take [conservatively] $r=\min(n_1, n_2)$ or use a more sophisticated formula.
[Note that our model does specify equal variances and so we could use the ‘pooled variance version’ of the t-test

\[
\frac{\bar{T}_1 - \bar{T}_2}{\sqrt{\text{var}(\bar{T}_1 - \bar{T}_2)}} \sim t_{n_1+n_2-2}
\]

where \( \text{var}(\bar{T}_1 - \bar{T}_2) = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \) but it should make little difference in practice.

Ex 7.1 (continued)

\[
\begin{array}{ccc}
B & \rightarrow & A \\
A & \rightarrow & B \\
n_i & 55 & 54 \\
\bar{T}_i & 176.28 & 180.17 \\
s_i & 26.56 & 26.27 \\
\end{array}
\]

so \( t = \frac{180.17 - 176.28}{\sqrt{\frac{26.27^2}{54} + \frac{26.56^2}{55}}} = 0.769 \) which is clearly non-significant when compared with \( t_{54} \) and so the data provide no evidence of a carry-over effect.

NB ‘pooled’ 2-sample \( t = \frac{180.17 - 176.28}{\sqrt{\frac{54 \cdot 26.27^2}{107} + \frac{53 \cdot 26.56^2}{107}}} \times \left( \frac{1}{55} + \frac{1}{54} \right) = 0.769 \)

(little difference because the variances are almost equal anyway)
7.3.1.1 Notes

- Test for carryover typically has low power since it involves **between** patient comparisons.

- If there is a significant carryover effect (i.e. treatment x period interaction) then it is **NOT SENSIBLE** to test for period and treatment separately, so
  a) plot out means and inspect
  b) just use first period results and compare A and B as a parallel group study.

- If just first period results are used then the treatment comparison is **between** patients (so also of low power).

- If there is a carryover then it means that the results of the second period are ‘contaminated’ and give no useful information on treatment comparisons — the trial should have been designed with a longer washout period.

- **NB** we used the *average* of the two values for each patient (i.e. from period 1 and period 2) in describing the carryover test since then the model indicates this has a mean of $\mu$ when there is no carryover. The value of the t-statistic would be **exactly** the same if we used just the sum of the two period values — this is easier (avoids dividing by 2!) and this will be the procedure in later examples.
7.3.2 Treatment & period effects

Consider \( D_{ik} = Y_{i1k} - Y_{i2k} \) i.e. within subject differences.

Then \( D_{1k} \sim N((\tau_A - \tau_B) + (\pi_1 - \pi_2), 2\sigma^2) \) group 1 and

\[ D_{2k} \sim N((\tau_B - \tau_A) + (\pi_1 - \pi_2), 2\sigma^2) \] group 2

7.3.2.1 Treatment test

\( H_0: \tau_A = 0 = \tau_B \)

If this is true, then \( D_{1k} \) and \( D_{2k} \) have identical distributions so we can test \( H_0 \) by a \( t \)-test for equality of means as before.

\[ \frac{\bar{D}_1 - \bar{D}_2}{\sqrt{\frac{s^2_{D1}}{n_1} + \frac{s^2_{D2}}{n_2}}} \sim t_r \]

where now \( s^2_{D1} \) is the sample variance of the differences \( D_{1k} \).

Notice that \( \bar{D}_1 \) is the difference between period 1 and period 2 results averaged over those in group 1 and \( \bar{D}_2 \) is the difference between period 1 and period 2 results averaged over those in group 2. Thus this test can be regarded as a two-sample \( t \)-test on period 1 – period 2 differences between the two groups of subjects.
Ex 7.1 (continued again)

\[ \begin{array}{c|c|c}
  & B \rightarrow A & A \rightarrow B \\
  n_i & 55 & 54 \\
  \bar{D}_i & 5.04 & -2.81 \\
  s_i & 15.32 & 19.52 \\
\end{array} \]

We have 
\[ t = \frac{5.04 - (-2.81)}{\sqrt{\frac{15.32^2}{55} + \frac{19.52^2}{54}}} = 2.33 \]

so \( p=0.024 \) when compared with \( t_{54} \) — significant evidence of treatment effects.

[The pooled t-statistic is 
\[ t = \frac{5.04 - (-2.81)}{\sqrt{\frac{54 \cdot 15.32^2 + 53 \cdot 19.52^2}{107}} \times \left( \frac{1}{55} + \frac{1}{54} \right)} = 2.34 \text{ with a } p\text{-value of 0.021 when compared with } t_{107} \text{ (i.e. no material or practical difference)}]
7.3.2.2 Period test

\[ H_0: \pi_1 = 0 = \pi_2 \]

If \( H_0 \) is true then \( D_{1k} \) and \( -D_{2k} \) will have identical distributions and so the test will be based on

\[
\frac{\bar{D}_1 - (-\bar{D}_2)}{\sqrt{\frac{s_{D1}^2}{n_1} + \frac{s_{D2}^2}{n_2}}} \sim t_r
\]

NB it is + in the numerator (not −) since it is still a 2-sample t-test of 2 sets of numbers the \( \{(Y_{11k} - Y_{12k}); k=1,…,n_1\} \) from group 1 and the \( \{(Y_{21k} - Y_{22k}); k=1,…,n_2\} \) from group 2.

Notice that \( \bar{D}_1 \) is the difference between \( Treatment \ A \) and \( Treatment \ B \) results averaged over those in group 1 and \( -(\bar{D}_2) \) is the difference between \( Treatment \ A \) and \( Treatment \ B \) results averaged over those in group 2. Thus this test can be regarded as a two-sample t-test on \( Treatment \ A - Treatment \ B \) differences between the two groups of subjects.

Ex 7.1 (continued yet again)

We have \( t = \frac{5.04 - (+2.81)}{3.365} = 0.66 \)

so no significant evidence of a period effect.

[Same conclusion from the pooled test]
7.4* Analysis with Linear Models

7.4.0* Introduction

The analyses presented above using carefully chosen t-tests provide an illustration of the careful use of an underlying model in selecting appropriate tests to examine hypotheses of interest. However, to extend the ideas to more complicated cross-over trails with more treatments and periods it is necessary to use a more refined analysis with linear models. The basic model for a multi-period multi-treatment trial for the response of patient k to treatment i in period j is:

\[ Y_{ijk} = \mu + \tau_i + \pi_j + \lambda_{ij} + \alpha_k + \epsilon_{ijk} \]

where \( \epsilon_{ijk} \sim N(0, \sigma^2) \), \( \alpha_k \sim N(0, \phi^2) \), \( \Sigma \alpha_i = \Sigma \tau_j = \Sigma \lambda_{ij} = 0 \) and where \( \lambda_{ij} \) denotes the carryover effect which mathematically is identical to an interaction between the factors treatment and period. Note that this model is slightly different from that given in §7.3 where the suffix i was used to indicate which group a patient belonged to and here it denotes the treatment received. The essence of a cross-over trial is that not all combinations of i, j and k are tested. For example in a trial with two periods and two treatments only about half of the patients will receive treatment 1 in period 1 and for others the combination i = j = 1 will not be used. Since the patient effect \( \alpha_k \) is specified as a random variable this is strictly a random effects model which is a topic covered in the second semester in MAS473/6003 so we present first an approximate analysis with a fixed effects model which alters the assumption that the \( \alpha_k \) are random variables and instead have the identifiability constraint \( \Sigma \alpha_k = 0 \).
7.4.1* Fixed effects analysis

The data structure presumed is that the dataframe consists of variable response with factors treatment, period and patient. Dataframes provided in the example data sets with this course are generally not in this form. Typically, in the example data sets the responses in the two periods are given as separate variables so each record consists of responses to one subject, which is convenient for performing the two sample t-tests described in earlier sections and these will require some manipulation.

The R analysis is then provided by:

```r
> crossfixed<-lm(result ~ period + treatment + patient + treatment:period)
> anova(crossfixed)
```

This will give an analysis of variance with entries for testing with F-tests differences between periods, treatments and the carryover (i.e. treatment×period interaction). The p-values will be almost the same as those from the separate t-tests and will be identical if non-default pooled variance t-tests are used by including `var.equal = TRUE` in the `t.test(.)` command.

Strictly speaking it has been presumed here that the numbers of subjects allocated to the various groups receiving treatments in the various orders have ensured that the factors period and treatment are orthogonal (e.g. equal number to two groups in a 2 periods 2 treatments trial). If this is not the case then the above analysis of variance will give a ‘periods ignoring treatments’ sum of squares and a ‘treatments adjusted for periods’ sum of squares. This aspect of the analysis may be discussed more fully in the second semester course MAS370/6012 or in MAS363/463/473/6003).
7.4.2* Random effects analysis

The same data structure is used and here the library \texttt{nlme} for random effects analysis is required and a random effects linear model is fitted with \texttt{lme(.)}

The R analysis is then provided by:

```r
> library(nlme)
> crossrandom<- lme(result ~ period + treatment + treatment:period, random = ~ 1|patient)
> anova(crossrandom)
```

The analysis of variance table will usually be very similar to that provided by the fixed effects model except that the standard errors of estimated parameters will be a little larger (to allow for the additional randomness introduced by regarding the patients as randomly selected from a broader population) and consequently the p-values associated with the various fixed effects of treatment, period and interaction will be a little larger (i.e. less significant).

7.4.3* Deferment of example

An example is not provided here but analyses using the two forms of model will be given on the hours sleep data used in Q2 on Task Sheet 4.
7.5 Notes

♦ If there is a substantial period effect, then it may be difficult to interpret any overall treatment difference within patients, since the observed treatment difference in any patient depends so much on which treatment was given first.

♦ Some authors (e.g. Senn, 2002) strongly disagree with the advisability of performing carryover tests. In part, the argument is based upon the difficulty introduced by a two-stage analysis, i.e. where the result of the first stage (a test for carryover) determines the form of the analysis for the second stage (i.e. whether data from both periods or just the first is used). This causes severe inferential problems since strictly the second stage is conditional upon the outcome of the first. In practice, most pharmaceutical companies rely upon medical considerations to eliminate the possibility of any carryover of treatments. In any case, the test for carryover typically has low power needs to be supplemented by medical knowledge — i.e. need expert opinion that either the two treatments cannot interact or that the washout period is sufficient, cannot rely purely on statistical evidence.
We can obtain confidence intervals for treatment differences since \( \frac{1}{2}(\bar{D}_1 - \bar{D}_2) \sim N(\tau_A - \tau_B, \frac{1}{2}\sigma^2(n_1^{-1} + n_2^{-1})) \) and estimate \( \sigma^2 \) with a pooled variance estimate or else say that the standard error of \( \frac{1}{2}(\bar{D}_1 - \bar{D}_2) \) is \( \sqrt{\frac{1}{4}\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)} \) and use the approximate formula for [say] a 95% CI of \( \frac{1}{2}(\bar{D}_1 - \bar{D}_2) \pm 2\times s.e.\{\frac{1}{2}(\bar{D}_1 - \bar{D}_2)\} \) (2 rather than 1.96 is adequate given the approximations made anyway in assuming normality etc).

If it is unsafe to assume normality the various two-sample t-tests above can be replaced by non-parametric equivalents, e.g. a Wilcoxon-Mann-Whitney test.

The simpler non-parametric test, a sign test, is essentially identical to the case of binary responses considered in §7.4 below.

Sample size & efficiency of crossover trials:

it can be shown that the number of patients required in a crossover trial is \( N = n(1 - \rho) \) where \( n= \) number required in each arm of a parallel group study and \( \rho= \) correlation between the 2 measurements on each patient (assuming no carryover effect). Since \( \rho > 0 \) usually, need fewer patients in a crossover than in a parallel group study. Sample size calculation facilities for cross-over trials are available in power.exe.
Can be extended to > 2 treatments and periods, usually when intervals between treatments can be very short.

\[
\begin{array}{ccc}
1 & 2 & 3 \\
A & B & B \\
B & A & A \\
A & B & C \\
C & A & B \\
B & C & A \\
\end{array}
\]

In trials involving several treatments it is unrealistic to consider all possible orderings and so need ideas of *incomplete block designs* [balanced or partially balanced] to consider a balanced subset of orderings. (See MAS370 or MAS6011 second semester).

Crossover trials are most suitable for short acting treatments where carryover effect is not likely, but usually not curative so baseline is similar in period 2.
7.6 Binary Responses
The analysis of binary responses introduces some new features but is essentially identical in logic to that of continuous responses considered above. The key idea is to consider within subject comparisons as before. This is achieved by considering whether the difference between the responses to the two treatments for the same subject indicates treatment A is ‘better’ or ‘worse’ than treatment B. If the responses on the two treatments are identical then that subject provides essentially no information on treatment differences.

7.6.1 Example: (Senn, 2002)
A two-period double blind crossover trial of 12μg formoterol solution compared with 200μg salbutamol solution administered to 24 children with exercise induced asthma. Response is coded as + and – corresponding to ‘good’ and ‘not good’ based upon the investigators overall assessment. Subjects were randomised to one of two groups: group 1 received the treatments in the order formoterol → salbutamol; group 2 in the order salbutamol → formoterol. The results are given below:
<table>
<thead>
<tr>
<th>group</th>
<th>subject</th>
<th>formoterol</th>
<th>salbutamol</th>
<th>preference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>—</td>
</tr>
<tr>
<td>group 1</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>+</td>
<td>+</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td>group 2</td>
<td>13</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>+</td>
<td>+</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>+</td>
<td>+</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>+</td>
<td>+</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>+</td>
<td>+</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>+</td>
<td>+</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>—</td>
<td>+</td>
<td>s</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>+</td>
<td>—</td>
<td>f</td>
</tr>
</tbody>
</table>
To test for a difference between treatments we test whether the proportion of subjects preferring the *first period treatment* is associated with which order the treatments are given in, (c.f. performing a two sample t-test on the period 1 – period 2 responses). This test is sometimes known as the Mainland-Gart Test:

<table>
<thead>
<tr>
<th>preference</th>
<th>sequence</th>
<th>first period</th>
<th>second period</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>for → sal</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>sal → for</td>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>10</td>
<td>6</td>
<td>16</td>
</tr>
</tbody>
</table>

The value of the Pearson chi-squared test statistic is

\[
(9 \times 6 - 1 \times 0)^2 \times 16 / [10 \times 6 \times 7 \times 9] = 12.34
\]

which is clearly significant at a level <0.001 and so the data provide strong evidence of superiority of the treatment by formoterol.

It might be noted here that the entries in this table are rather small. More relevantly, the *expected values* of the cell values are small with two of the less than 5. This means that the chi-squared distribution is not an adequate approximation to the null distribution of the test statistic and so in calculating the p-value we either need to simulate the p-value or use a Fisher exact test:

```r
x <- matrix(c(9, 0, 1, 6), ncol=2)
chisq.test(x, simulate.p.value=T, B=1000000)$p.value
fisher.test(x)$p.value
```
To test for a period effect we similarly test whether the proportion of subjects preferring treatment A is associated with the order in which the treatments are given:

<table>
<thead>
<tr>
<th>sequence</th>
<th>formoterol</th>
<th>salbutamol</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>for → sal</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>sal → for</td>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>total</td>
<td>15</td>
<td>1</td>
<td>16</td>
</tr>
</tbody>
</table>

Now the test statistic is $(9 	imes 1 - 6 	imes 0)^2 	imes 16 / [15 	imes 1 	imes 7 	imes 9] = 1.37$ and we conclude that there is no evidence of a period effect.

### 7.7 Summary and Conclusions

Possible effects that must be tested in a two-treatment two-period crossover trial (whether continuous or binary outcomes) are:

- **carryover:** test by two-sample test on average response over both periods
- **treatment:** test by two-sample test on differences of period I – period II results between the two groups of subjects
- **period:** test by two-sample test on differences of treatment A – treatment B results between the two groups of subjects.
If carryover (i.e. treatment×period interaction) is present then use only results from period I, in which case treatment comparisons are between subjects. A full crossover analysis gives a within subject comparison.

♦ Use of a preliminary test for carryover is not recommended by some authorities and it is preferable to rely upon medical considerations to eliminate the possibility of a carryover.

♦ If normality is assumed then the tests can be performed with two sample t-tests. These can be replaced with non-parametric equivalents such as a Wilcoxon-Mann-Whitney test.

♦ binary responses can be analyzed with a Mainland-Gart test which considers only those subjects exhibiting different responses to the treatments.
Tasks 5

1) Senn and Auclair (Statistics in Medicine, 1990, 9) report on the results of a clinical trial to compare the effects of single inhaled doses of 200\,\mu g salbutamol (a well established bronchodilator) and 12\,\mu g formoterol (a more recently developed bronchodilator) for children with moderate or severe asthma. A two-treatment, two-period crossover design was used with 13 children entering the trial, and the observations of the peak expiratory flow, a measure of lung function where large values are associated with good responses, were taken. The following summary of the data is provided.

<table>
<thead>
<tr>
<th>Group 1: formoterol → salbutamol (n₁ = 7)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Period 1</td>
<td>337.1</td>
<td>306.4</td>
<td>643.6</td>
<td>30.7</td>
</tr>
<tr>
<td>Period 2</td>
<td>53.8</td>
<td>64.7</td>
<td>114.3</td>
<td>33.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group 2: salbutamol → formoterol (n₂ = 6)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Period 1</td>
<td>283.3</td>
<td>345.8</td>
<td>629.2</td>
<td>-62.6</td>
</tr>
<tr>
<td>Period 2</td>
<td>105.4</td>
<td>70.9</td>
<td>174.0</td>
<td>44.7</td>
</tr>
</tbody>
</table>

a) Specify a model for peak expiratory flow which incorporates treatment, period and carryover effects.

b) Assess the carryover effect, and, if appropriate, investigate treatment differences. In each case specify the hypotheses of interest and illustrate the appropriateness of the test.
2) A and B are two hypnosis treatments given to insomniacs one week apart. The order of receiving the treatment is randomized between patients. The measured response is the number of hours sleep during the night. Data are given in the following table.

<table>
<thead>
<tr>
<th>patient</th>
<th>period 1</th>
<th>period 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>B</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
<td>3</td>
</tr>
</tbody>
</table>

a) Calculate the mean for each treatment in each period and display the results graphically.
b) Assess the carryover effect.
c) If appropriate, assess the treatment and period effects.
Exercises 3

1) Given below is an edited extract from an SPSS session analysing the results of a two period crossover trial to investigate the effects of two treatments A (standard) and B (new) for cirrhosis of the liver. The figures represent the maximal rate of urea synthesis over a short period and high values are desirable. Patients were randomly allocated to two groups: the 8 subjects in group 1 received treatment A in period 1 and B in period 2. Group 2 (13 subjects) received the treatments in the opposite order.
   i) Specify a suitable model for these data which incorporates treatment, period and carryover effects.
   ii) Assess the evidence that there is a carryover effect from period 1 to period 2.
   iii) Do the data provide evidence that there is a difference in average response between periods 1 and 2?
   iv) Assess whether the treatments differ in effect, taking into account the results of your assessments of carryover and period effects.
   v) Repeat the statistical analysis in R
   vi) ★The final stage in the analysis recorded below produced 95% Confidence Intervals, firstly, for the mean differences in response between periods 1 and 2 for the 21 subjects and, secondly, for the mean differences in response to treatments A and B for the 21 subjects. By referring to your model for these data, explain why these two confidence intervals can not be used to provide indirect tests of the hypotheses of no period and no treatment effects respectively.
### Extract from SPSS Analysis of Crossover Trial on Liver Treatment

#### Summarize

<table>
<thead>
<tr>
<th>Patnum</th>
<th>Group</th>
<th>Period1</th>
<th>Period2</th>
<th>Sum1+2</th>
<th>PeriodDiff</th>
<th>TreatDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>48.00</td>
<td>51.00</td>
<td>99.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>2.00</td>
<td>1.00</td>
<td>43.00</td>
<td>47.00</td>
<td>90.00</td>
<td>-4.00</td>
<td>-4.00</td>
</tr>
<tr>
<td>3.00</td>
<td>1.00</td>
<td>60.00</td>
<td>66.00</td>
<td>126.00</td>
<td>-6.00</td>
<td>-6.00</td>
</tr>
<tr>
<td>4.00</td>
<td>1.00</td>
<td>35.00</td>
<td>40.00</td>
<td>75.00</td>
<td>-5.00</td>
<td>-5.00</td>
</tr>
<tr>
<td>5.00</td>
<td>1.00</td>
<td>36.00</td>
<td>39.00</td>
<td>75.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>6.00</td>
<td>1.00</td>
<td>43.00</td>
<td>46.00</td>
<td>89.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>7.00</td>
<td>1.00</td>
<td>46.00</td>
<td>52.00</td>
<td>98.00</td>
<td>-6.00</td>
<td>-6.00</td>
</tr>
<tr>
<td>8.00</td>
<td>1.00</td>
<td>54.00</td>
<td>42.00</td>
<td>96.00</td>
<td>12.00</td>
<td>12.00</td>
</tr>
<tr>
<td>9.00</td>
<td>2.00</td>
<td>31.00</td>
<td>34.00</td>
<td>65.00</td>
<td>-3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>10.00</td>
<td>2.00</td>
<td>51.00</td>
<td>40.00</td>
<td>91.00</td>
<td>11.00</td>
<td>-11.00</td>
</tr>
<tr>
<td>11.00</td>
<td>2.00</td>
<td>31.00</td>
<td>34.00</td>
<td>65.00</td>
<td>-3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>12.00</td>
<td>2.00</td>
<td>43.00</td>
<td>36.00</td>
<td>79.00</td>
<td>7.00</td>
<td>-7.00</td>
</tr>
<tr>
<td>13.00</td>
<td>2.00</td>
<td>47.00</td>
<td>38.00</td>
<td>85.00</td>
<td>9.00</td>
<td>-9.00</td>
</tr>
<tr>
<td>14.00</td>
<td>2.00</td>
<td>29.00</td>
<td>32.00</td>
<td>61.00</td>
<td>-3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>15.00</td>
<td>2.00</td>
<td>35.00</td>
<td>44.00</td>
<td>79.00</td>
<td>-9.00</td>
<td>9.00</td>
</tr>
<tr>
<td>16.00</td>
<td>2.00</td>
<td>58.00</td>
<td>50.00</td>
<td>108.00</td>
<td>8.00</td>
<td>-8.00</td>
</tr>
<tr>
<td>17.00</td>
<td>2.00</td>
<td>60.00</td>
<td>60.00</td>
<td>120.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>18.00</td>
<td>2.00</td>
<td>82.00</td>
<td>63.00</td>
<td>145.00</td>
<td>19.00</td>
<td>-19.00</td>
</tr>
<tr>
<td>19.00</td>
<td>2.00</td>
<td>51.00</td>
<td>50.00</td>
<td>101.00</td>
<td>1.00</td>
<td>-1.00</td>
</tr>
<tr>
<td>20.00</td>
<td>2.00</td>
<td>49.00</td>
<td>42.00</td>
<td>91.00</td>
<td>7.00</td>
<td>-7.00</td>
</tr>
<tr>
<td>21.00</td>
<td>2.00</td>
<td>47.00</td>
<td>43.00</td>
<td>90.00</td>
<td>4.00</td>
<td>-4.00</td>
</tr>
</tbody>
</table>

#### T-Test

<table>
<thead>
<tr>
<th></th>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>t</th>
<th>Df</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum1+2</td>
<td>2.7308</td>
<td>8.7046</td>
<td>.314</td>
<td>18.683</td>
<td>.757</td>
</tr>
<tr>
<td>PeriodDiff</td>
<td>-5.9423</td>
<td>2.9429</td>
<td>-2.019</td>
<td>17.646</td>
<td>.059</td>
</tr>
<tr>
<td>TreatDiff</td>
<td>1.4423</td>
<td>2.9429</td>
<td>.490</td>
<td>17.646</td>
<td>.630</td>
</tr>
</tbody>
</table>
## Summarize

Case Summaries(a)

<table>
<thead>
<tr>
<th>GROUP</th>
<th>Summ1+2</th>
<th>PeriodDiff</th>
<th>TreatDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>99.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>2</td>
<td>90.00</td>
<td>-4.00</td>
<td>-4.00</td>
</tr>
<tr>
<td>3</td>
<td>126.00</td>
<td>-6.00</td>
<td>-6.00</td>
</tr>
<tr>
<td>4</td>
<td>75.00</td>
<td>-5.00</td>
<td>-5.00</td>
</tr>
<tr>
<td>5</td>
<td>75.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>6</td>
<td>89.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>7</td>
<td>98.00</td>
<td>-6.00</td>
<td>-6.00</td>
</tr>
<tr>
<td>8</td>
<td>96.00</td>
<td>12.00</td>
<td>12.00</td>
</tr>
<tr>
<td>Total</td>
<td>N</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>93.5000</td>
<td>-2.2500</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>16.1688</td>
<td>5.8979</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GROUP</th>
<th>Summ1+2</th>
<th>PeriodDiff</th>
<th>TreatDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>65.00</td>
<td>-3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>2</td>
<td>91.00</td>
<td>11.00</td>
<td>-11.00</td>
</tr>
<tr>
<td>3</td>
<td>65.00</td>
<td>-3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>4</td>
<td>79.00</td>
<td>7.00</td>
<td>-7.00</td>
</tr>
<tr>
<td>5</td>
<td>85.00</td>
<td>9.00</td>
<td>-9.00</td>
</tr>
<tr>
<td>6</td>
<td>61.00</td>
<td>-3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>7</td>
<td>79.00</td>
<td>-9.00</td>
<td>9.00</td>
</tr>
<tr>
<td>8</td>
<td>108.00</td>
<td>8.00</td>
<td>-8.00</td>
</tr>
<tr>
<td>9</td>
<td>120.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>145.00</td>
<td>19.00</td>
<td>-19.00</td>
</tr>
<tr>
<td>11</td>
<td>101.00</td>
<td>1.00</td>
<td>-1.00</td>
</tr>
<tr>
<td>12</td>
<td>91.00</td>
<td>7.00</td>
<td>-7.00</td>
</tr>
<tr>
<td>13</td>
<td>90.00</td>
<td>4.00</td>
<td>-4.00</td>
</tr>
<tr>
<td>Total</td>
<td>N</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>90.7692</td>
<td>3.6923</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>23.6684</td>
<td>7.4876</td>
</tr>
</tbody>
</table>

| Total | N       | 21         | 21        |
|       | Mean    | 91.8095    | 1.4286    |
|       | Std. Deviation | 20.7235 | 7.3863 |

## Explore

<table>
<thead>
<tr>
<th>PeriodDiff</th>
<th>95% Confidence Interval for Mean</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95% Confidence Interval for Mean</td>
<td>-1.9336</td>
<td>4.7908</td>
</tr>
</tbody>
</table>

© NRJF, 1996 ➔
| TreatDiff | 95% Confidence Interval for Mean | -6.2411 | -0.044571 |
8. Combining trials

8.1 Small trials

Some trials are too small to have much chance of picking up differences when they exist (perhaps because of insufficient care over power and sample size)

Problem 1:—

Non-significant test result interpreted by clinicians as ‘two treatments are the same’ even though the test may have been so low in power that it was not able to detect a real difference

Problem 2:—

Small trials giving non-significant results are hardly ever published: publication bias — medical literature contains all large trials and the significant small trials.

Solutions

a) do not publish any small trials

b) combine trials
8.2 Pooling trials and meta analysis

We may have results from several trials or centres. How should we combine them?

e.g. For a binary response of treatment vs placebo
e.g. trial j (for j=1,2,......,N)

\[
\begin{align*}
&\text{Treatments} & & \text{Successes} & & \text{Failures} \\
&\text{Y}_{1j} & & n_{1j} - Y_{1j} & & n_{1j} \\
&\text{Y}_{2j} & & n_{2j} - Y_{2j} & & n_{2j} \\
&t_j & & n_j - t_j & & n_j
\end{align*}
\]

It can be dangerous to collapse these \( N \times 2 \) separate tables into 1 single \( 2 \times 2 \) table:

\[
\begin{array}{c|cc|}
\text{centre 1} & \text{S} & \text{F} & 30\%\text{S} \\
\hline
\text{trt} & 30 & 70 & \\
\text{plac} & 120 & 180 & 40\%\text{S} \\
\hline
\text{total} & 150 & 250 & \\
\end{array}
\quad
\begin{array}{c|cc|}
\text{centre 2} & \text{S} & \text{F} & 70\%\text{S} \\
\hline
\text{trt} & 210 & 90 & 70\%\text{S} \\
\text{plac} & 80 & 20 & 80\%\text{S} \\
\hline
\text{total} & 290 & 110 & \\
\end{array}
\]

looks like \textbf{placebo} better? \( (\chi^2 = 3.2, \text{n.s.}) \)

looks like \textbf{placebo} better? \( (\chi^2 = 3.76, \text{n.s.}) \)
but if we collapse the two tables into one:

<table>
<thead>
<tr>
<th></th>
<th>centre 1 &amp; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>S</td>
</tr>
<tr>
<td>240</td>
<td>160</td>
</tr>
<tr>
<td>plac</td>
<td>200</td>
</tr>
<tr>
<td>440</td>
<td>360</td>
</tr>
</tbody>
</table>

This is known as **Simpson’s Paradox** — it is misleading to look at margins of higher dimensional arrays, especially when there are imbalances in treatment numbers or in the magnitudes of the effects.

The root cause of the paradox here is that the overall success rates in the two centres is markedly different (30–40% in centre 1 but 70–80% in centre 2) so it is misleading to ignore the **centre differences** and add the results together from them.
8.3 Mantel-Haenszel Test

One way of combining data from such trials is using the Mantel-Haenszel test (but this does not necessarily overcome Simpson’s Paradox — it only avoids differences BETWEEN trials and assesses evidence WITHIN trials).

Consider a single 2×2 table:

<table>
<thead>
<tr>
<th>Treatments</th>
<th>Successes</th>
<th>Failures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y₁</td>
<td>n₁−Y₁</td>
<td>n₁</td>
</tr>
<tr>
<td>Y₂</td>
<td>n₂−Y₂</td>
<td>n₂</td>
</tr>
<tr>
<td>t</td>
<td>n−t</td>
<td>n</td>
</tr>
</tbody>
</table>

and assume Yᵢ ~ B(nᵢ,θᵢ) ; i=1,2

interested in H₀: θ₁ = θ₂

Fisher’s exact test considers

P(y₁,y₂|y₁+y₂=t) i.e. conditions on the total number of successes

If θ₁ = θ₂ then P(y₁,y₂|y₁+y₂=t) = \[ \frac{\binom{n₁}{y₁}\binom{n₂}{t−y₁}}{\binom{n}{t}} \]

(i.e. a hypergeometric probability)

⇒ E(Y₁)=n₁t/n and V(Y₁)=n₁n₂t(n−t)/n²(n−1)
So, if we have large margins, a means of analysis is to say that

\[ T_{MH} = \frac{[Y_1-E(Y_1)]^2}{V(Y_1)} \sim \chi^2_1 \text{ under } H_0 \]

If \( T_{MH} > \chi^2_{1.1-\alpha} \) then \( p < \alpha \) and there is a significant treatment difference.

8.3.1 Comments

1. Asymptotically equivalent to usual \( \chi^2 \) test.

2. Known as the Mantel-Haenszel [or very misleadingly as a Randomization test].

3. Does not matter whether you use \( Y_1, Y_2, n-Y_1 \) or \( n-Y_2 \).

4. The extension to several tables is simple. Keeping the \( k \) tables separate we calculate \( E(Y_{1j}) \) and \( \text{var}(Y_{1j}) \) from each of the tables, \( j=1,...,k \). We use \( W=\Sigma Y_{1j} \) and under \( H_0: \theta_1 = \theta_2 \) in each table, i.e. \( \theta_{1j} = \theta_{2j} \), i.e. response ratio equal within each study we have \( E(W) = \Sigma E(Y_{1j}) \) and \( V(W) = \Sigma V(Y_{1j}) \) and \( [W-E(W)]^2/V(W) \sim \chi^2_1 \text{ under } H_0 \) again.

5. This test is most appropriate when treatment differences are consistent across tables (we can test this but it is easier in a logistic regression framework — see later) — the test pools evidence from within the different trials whilst avoiding differences between trials.
8.3.2 Possible limitations of M-H test

- Randomness dubious
- Reporting bias
- Not clear that $\theta_i$ is the same for all trials.

8.3.3 Relative merits of M-H & Logistic Regression approaches

The Mantel-Haenszel test is simpler if one has just 2 qualitative prognostic factors to adjust for and wishes only to assess significance, not magnitude, of a treatment difference. The logistic approach (see below) is more general and can include other covariates, further, it can test whether treatment differences are consistent across tables. The M-H test is not very appropriate for assessing effects if tables are inhomogeneous, i.e. if treatment differences are inconsistent across tables, and must be used with care if success rates differ markedly (i.e. leading to Simpson’s Paradox).
8.3.4 Example: pooling trials

A research worker in a skin clinic believes that the severity of eczema in early adulthood may depend on breast or bottle feeding in infanthyhood and that bottle fed babies are more likely to suffer more severely in adulthood. Sufferers of eczema may be classified as ‘severe’ or ‘mild’ cases. The research worker finds that in a random sample of 20 cases in his clinic who were bottle fed, 16 were ‘severe’ whilst for 20 breast fed cases only 10 were ‘severe’. How do you assess the research workers belief?

In a search through the recent medical literature he finds the results, shown below, of two more extensive studies which have been carried out to investigate the same question. Assess the research worker's belief in the light of the evidence from these studies.

<table>
<thead>
<tr>
<th>study</th>
<th>Bottle fed</th>
<th>Breast fed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>severe</td>
<td>mild</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>34</td>
</tr>
</tbody>
</table>
Analysis

Study 1

<table>
<thead>
<tr>
<th></th>
<th>Severe</th>
<th>Mild</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottle</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>Breast</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>14</td>
</tr>
</tbody>
</table>

\[ Y_1 = \text{number of response 'severe' on bottle fed.} \]

Under \( H_0 \) response ratios equal:

\[
\begin{align*}
E(Y_1) &= 20 \times 26 / 40 = 13 \\
V(Y_1) &= 20 \times 20 \times 26 \times 14 / 40 \times 40 \times 39 = 2.333
\end{align*}
\]

So Mantel-Haenszel test statistic is

\[
\frac{(16-13)^2}{2.333} = 3.86 > \chi^2_{0.95} = 3.84
\]

and so is just significant at 5% level, i.e. more severe cases on bottle feed.
Study 2

<table>
<thead>
<tr>
<th></th>
<th>Severe</th>
<th>Mild</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottle</td>
<td>34</td>
<td>16</td>
<td>50</td>
</tr>
<tr>
<td>Breast</td>
<td>30</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>36</td>
<td>100</td>
</tr>
</tbody>
</table>

\[ E(Y_2) = \frac{50 \times 64}{100} = 32 \]
\[ V(Y_2) = 5.8182 \]

M-H test statistic = 0.687, \( p > 0.05 \), n.s.

Study 3

<table>
<thead>
<tr>
<th></th>
<th>Severe</th>
<th>Mild</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottle</td>
<td>80</td>
<td>34</td>
<td>114</td>
</tr>
<tr>
<td>Breast</td>
<td>48</td>
<td>50</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>84</td>
<td>212</td>
</tr>
</tbody>
</table>

\[ E(Y_3) = 68.83, \quad V(Y_3) = 12.6668, \]

M-H test statistic = 9.850, \( p < 0.005 \)
Combining all 3 studies

Use \( W = Y_1 + Y_2 + Y_3 \).

Under \( H_0 \): response ratios equal,

\[
W = 130, \quad E(W) = 113.83, \quad V(W) = 20.8183 \quad \text{so}
\]

M-H test statistic = 12.56, \( p < 0.0005 \), highly significant

**Caution:** the response ratios in the three studies differ quite a lot

(80\%, 68\% and 70\% in studies 1, 2 and 3)

For interest, combining all 3 tables gives:

<table>
<thead>
<tr>
<th></th>
<th>Severe</th>
<th>Mild</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottle</td>
<td>130</td>
<td>54</td>
</tr>
<tr>
<td>Breast</td>
<td>88</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>218</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td></td>
<td>352</td>
</tr>
</tbody>
</table>

giving an Pearson \( \chi^2 \)–statistic of 12.435, \( p < 0.0005 \). It might also be noted that the M-H statistic calculated from this table is slightly different, 12.400. These small differences are inconsequential in this case. The combined M-H statistic tests for association *within strata*, i.e. within studies, and so avoids differences *between strata*, thus avoiding Simpson’s paradox (rather than overcoming it).

**Note:** We could also calculate the ordinary Pearson chi-squared values for each of these tables; the results are very close to (actually slightly greater than) the Mantel-Haenszel values since the numbers are large.
8.3.5 Example of Mantel-Haenszel Test in R

The function for performing a Mantel-Haenszel test in R is `mantelhaen.test()`. The Help system gives full details and examples.

The data are from the example 8.1 in §8.3.4 on page 135. The first example shews how to set up R to run a MH test on just one table by creating a factor \( z \) which has just one level.

```r
> x<-factor(rep(c(1,2),c(20,20)),labels=c("bottle","breast"))
> y<-factor(rep(c(1,2,1,2),c(16,4,10,10)),labels=c("severe","mild"))
> z<-factor(rep(1,40),labels="study 1")
> table(x,y,z)

, , study 1
    severe mild
bottle  16    4
breast  10   10
> mantelhaen.test(x,y,z,correct=F)

Mantel-Haenszel chi-square test without continuity correction

data:  x and y and z
Mantel-Haenszel chi-square = 3.8571, df = 1
, p-value = 0.0495

> 
```
The second example shews how to calculate the MH statistic for all three tables combined.

```R
> x<-factor(rep(c(1,2,1,2,1,2),c(20,20,50,50,114,98)),
+ labels=c("bottle","breast"))
> y<-factor(rep(c(1,2,1,2,1,2,1,2,1,2,1,2,1,2),
+ c(16,4,10,10,34,16,30,20,80,34,48,50)),
+ labels=c("severe","mild"))
> z<-factor(rep(c(1,2,3),c(40,100,212)),
+ labels=c("study 1","study 2","study 3"))
> table(x,y,z)

, , study 1
  severe mild
bottle  16   4
breast  10  10

, , study 2
  severe mild
bottle  34  16
breast  30  20

, , study 3
  severe mild
bottle  80  34
breast  48  50
> mantelhaen.test(x,y,z,correct=F)

Mantel-Haenszel chi-square test without continuity correction

data:  x and y and z
Mantel-Haenszel chi-square = 12.5593, df = 1, p-value = 0.0004
> 
```
8.4 Summary and Conclusions

♦ Combining trials can give paradoxical results if response rates and sample sizes are very different in the trials (Simpson’s Paradox)

♦ Simpson’s paradox can be resolved by more sophisticated modelling allowing for a separate ‘trial effect’

♦ The Mantel-Haenszel test provides an alternative way of analysing $2 \times 2$ tables which makes it easier to combine results from different trial but which does not overcome Simpson’s Paradox but avoids it.
Tasks 6

1) Two ointments A and B have been widely used for the treatment of athlete's foot. In a recent report the following results were noted, where response indicated temporary relief from the outbreak.

<table>
<thead>
<tr>
<th>Response</th>
<th>No Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ointment A</td>
<td>174</td>
</tr>
<tr>
<td>Ointment B</td>
<td>149</td>
</tr>
</tbody>
</table>

a) Based on these results the report concluded that ointment A was more effective than ointment B. Use the Mantel-Haenszel test to verify this conclusion.

b) Further investigation into the source of the data revealed that the data had been pooled from two clinics. The results from individual clinics were:

<table>
<thead>
<tr>
<th>Ointment A</th>
<th>Ointment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinic</td>
<td>Response</td>
</tr>
<tr>
<td>1</td>
<td>129</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
</tr>
</tbody>
</table>

Reassess the evidence in the light of these additional facts.
2) (Artificial data from Ben Goldacre, 06/08/11).

Imagine a study was conducted to examine the relationship between heavy drinking of alcohol and developing lung cancer, obtaining the following results:

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>No cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drinker</td>
<td>366</td>
<td>2300</td>
</tr>
<tr>
<td>Non-Drinker</td>
<td>98</td>
<td>1856</td>
</tr>
</tbody>
</table>

a) Calculate the ratio of the odds of developing cancer for drinkers to non-drinkers. What conclusions do you draw from this odds ratio?

b) It transpires that 330 of the drinkers developing cancer were smokers and 1100 of the drinkers who smoked did not, with corresponding figures for the non-drinkers of 47 and 156. Calculate the odds ratios separately for smokers and non-smokers. What conclusions do you draw?
9. Binary Response Data

9.1 Background

Responses are often measured on a binary or categorical scale. Here we only look at the binary case, so we can represent the response of the $i^{th}$ patient by $y_i = 1$ (success) or $y_i = 0$ (failure). We can use standard Pearson $\chi^2$ or Mantel-Haenszel tests but not all cross-classified tables are appropriate for application of these hypotheses tests of independence of classification or homogeneity. In some cases it is appropriate to consider different statistics calculated from the table to reflect on the key question of interest there are further techniques for special designs (e.g. paired observations) or if we have additional data, e.g. on covariates (such as different centres).

9.2 Observational Studies

9.2.1 Introduction

In epidemiological studies where it is not possible to control treatments or other factors administered to subjects inferences have to be based on observing characteristics and other events on subjects. For example, to investigate the effect of smoking on health (e.g. heart disease) cases of subjects with heart disease might be collected. These would be compared with controls who do not exhibit such symptoms but are otherwise similar to the cases in general respects (e.g. age, weight etc.) and the incidence of smoking in the two groups would be compared. This is an example of a retrospective study. A different form of observational study is a prospective study where a cohort of subjects who are known to have been exposed to some risk factor
(e.g. a very premature birth) and are followed up through a period. They are then observed at some later date and the incidence of a condition (e.g. school achievement very far below average) is assessed. In such studies the numbers of observations is typically very large since the incidence of the condition is often rare. It would be possible to use a chi-squared or a Mantel-Haenszel test for comparing the proportions but this would not be informative, either because with such large numbers of subjects the statistical test is very powerful and so return a highly significant result without saying anything about the magnitude of the effect or because the incidence is so rare that expected numbers in some cells are unduly low. Instead such observational studies are more traditionally analysed by estimating quantities that are of direct interpretability (odds ratios and relative risks) and they are assessed by calculating confidence intervals for their true values using formulae giving approximations to their standard errors.

9.2.2 Prospective Studies — Relative Risks

Prospective studies follow a group of subjects with different characteristics to see if an outcome of interest occurs. These would be used where the characteristic is not a ‘treatment’ that can be administered to a randomly selected group of subjects but some ‘risk factor’ such as very low birth weight or more than one month premature birth or blood group. The outcome may be some feature which occurs at some time later. The analysis would be based on calculating the risks of developing the feature for the different groups and, in the case of two outcomes (positive and negative say) and two groups (exposed and non-exposed say) calculating the relative risks.
The risk of a positive outcome for the exposed group is $a/(a+b)$ and for the non-exposed group it is $c/(c+d)$. The **relative risk** is the ratio of these two

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)}$$

and we compare this with the value 1 (the RR if there is no difference in risks for the two groups) by using its standard error.

The formula for the standard error of $\log_e(RR)$ is

$$S.E.\{\log_e(RR)\} = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$
9.2.2.1 Example

The data are taken from a study of ‘small-for-date’ babies who were classified as having symmetric or asymmetric growth retardation in relation to their Apgar score.

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symmetric</td>
<td>2</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Asymmetric</td>
<td>33</td>
<td>58</td>
<td>91</td>
</tr>
</tbody>
</table>

The calculations give RR = 0.3447, $\log_e(\text{RR}) = -1.0651$, 
s.e.$(\log_e(\text{RR})) = 0.6759$.

A 90% CI for $\log_e(\text{RR})$ is $-1.0651 \pm 1.645 \times 0.6759 = (-2.1769, 0.0467)$ and taking exponentials of this gives a 90% CI for the RR as $(0.11, 1.05)$. Since this interval contains 1 there is no evidence at the 10% level of a difference in risk of a low Apgar score between the two groups.
9.2.3 Retrospective Studies — Odds Ratios

Retrospective studies identify a collection of cases (e.g. with a disease) and compare these with respect to exposure to a risk factor with a group of controls (without the disease). The selection of the subjects is based on the outcome and not the characteristic defining the group as with prospective studies.

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Non-exposed</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

It is not sensible to calculate the risk of ‘being a case’ \( \frac{a}{a+b} \) since this can apparently be made any value just by selecting more or fewer controls which would increase or decrease \( b \) but not any other value.

Instead it is sensible to look at the odds of exposure for the cases and for the controls and look at the ratio between these. If exposure is not a risk factor for being a case then this odds ratio will be close to 1. As before there is a simple formula for the standard error of the \( \log_e \) of the odds ratio

\[
\text{OR} = \frac{a/c}{b/d} = \frac{ad}{bc}
\]

and

\[
\text{S.E.}(\log_e(\text{OR})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}
\]
9.2.3.1 Example

The following gives the results of a case-control study of erosion of dental enamel in relation to amount of swimming in a chlorinated pool.

<table>
<thead>
<tr>
<th>Swimming per week</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 6 hours</td>
<td>32</td>
<td>118</td>
</tr>
<tr>
<td>&lt; 6 hours</td>
<td>17</td>
<td>127</td>
</tr>
</tbody>
</table>

The calculations give OR=2.0259, s.e.(log_e(RR))=0.3262 and so a 95% for the log odds ratio is (0.0666, 1.3454) and the confidence interval for the odds ratio itself is thus (1.0689, 3.8397) which excludes the value 1 and so provides evidence at the 5% level of a raised risk of dental erosion in those swimming more than 6 hours a week.
9.3 Matched pairs

9.3.1 Introduction

In the comparison of two treatments A & B, suppose each patient receives both treatments (in random order), e.g. a crossover or matched-pair trial. We then observe pairs:

\[(y_{i1}, y_{i2})\]

response to A \quad response to B

of the form (0, 0), (0, 1), (0, 1), (1, 1), (1, 0), (1, 1), .......

E.g. Rheumatoid arthritis study, two treatments A & B.

Response caused? 1=yes, 0=no

Could present results as:

<table>
<thead>
<tr>
<th>response</th>
<th>yes</th>
<th>no</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>11</td>
<td>37</td>
<td>48</td>
</tr>
<tr>
<td>B</td>
<td>20</td>
<td>28</td>
<td>48</td>
</tr>
</tbody>
</table>

and then it is tempting to analyse this as

an ordinary 2×2 table with a \(\chi^2\)-test.

This \& INVALID \& since it ignores the double use of each patient (there are only 48 independent subjects in the table not 96).
A more useful summary is

\[
\begin{array}{ccc}
 & \text{yes} & \text{no} \\
\text{A} & \text{yes} & 8 & 3 & 11 \\
 & \text{no} & 12 & 25 & 37 \\
\text{B} & 20 & 28 & 48 \\
\end{array}
\]

A suitable test for what is really of interest (treatment difference) — not ‘no association’) is:

**9.3.2 McNemar’s Test**

Ignore (1,1) and (0,0), use the unlike pairs only. If no treatment differences exist, then the proportions of (1,0)’s (say) out of the total number of (1,0)’s and (0,1)’s should be consistent with binomial variation with \( p=\frac{1}{2} \).

In example
There are 3 (1,0)’s out of a total of 15 unlike pairs.

i.e. significance probability = \( 2 \times \sum_{x=0}^{3} \binom{15}{x} (\frac{1}{2})^{15-x} \) =0.035 which is significant at the 5% level.

For larger \( n \) use the Normal approximation

\[
\frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}} \sim \chi_1^2
\]

**Note:** We have not used the data from subjects where the responses were the same, i.e. subjects for whom both treatments
produced successes or both failures. This is sensible since these subjects provide no evidence on treatment differences, even though intuitively the results from these subjects might suggest that the two treatments are equivalent.
9.4 Logistic Modelling

9.4.1 Introduction

(for more details of logistic models see PAS372 or PAS6003)

Logistic modelling has become a very popular way of handling binary data and the analyses can be handled in most standard statistical packages.

In the clinical trials context define:

For patient $i$, outcome $Y_i = 0$ (failure) or 1 (success).

Treatment $x_i = 0$ (placebo) or 1 (treatment)

Then an alternative parameterization of the 2×2 set up is

$$P[Y_i = 1] = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \quad \text{and} \quad P[Y_i = 0] = 1 - P[Y_i = 1] = \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}$$

i.e. on placebo

$$P[Y_i = 1] = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

and on treatment $P[Y_i = 1] = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$

We can see that $\ln \left( \frac{P[Y_i = 1]}{P[Y_i = 0]} \right) = \beta_0 + \beta_1 x_i$
The model extends naturally to include other **prognostic factors** or **covariates**:

\[
\ln \left( \frac{P(Y_i = 1)}{P(Y_i = 0)} \right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip}
\]

\[= \beta_0 + \beta' x_i\]

where the \(x_{ij}\) can be continuous or discrete or dummy.

\[
\frac{P(Y_i = 1 \mid x_i)}{P(Y_i = 0 \mid x_i)} = \exp\{\beta_0 + \beta' x_i\}
\]

In this case \(P(Y_i=1) = P(\text{success}) = \frac{e^{\beta_0 + \beta' x}}{1 + e^{\beta_0 + \beta' x}} = \theta_i\)

and

\[
\ln \left( \frac{P(Y_i = 1)}{P(Y_i = 0)} \right) = \ln\left( \frac{\theta_i}{1 - \theta_i} \right) = \beta_0 + \beta' x_i
\]
9.4.2 Interpretation

For comparative trials

\[
\ln \left( \frac{P[Y_i = 1]}{P[Y_i = 0]} \right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_p X_{ip} \quad \text{if } x_{i1} = 0,
\]

i.e. on placebo

\[
\ln \left( \frac{P[Y_i = 1]}{P[Y_i = 0]} \right) = \beta_0 + \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_p X_{ip} \quad \text{if } x_{i1} = 1,
\]

i.e. on treatment

so if \( \beta_1 > 0 \), odds in favour of success are greater in treatment group and if \( \beta_1 < 0 \), odds in favour of success are greater in placebo group.

Similar interpretations for other factors:

\( \beta_j > 0 \Rightarrow P(\text{success}) \uparrow \text{ as } x_j \uparrow \text{ and } P(\text{success}) \downarrow \text{ as } x_j \downarrow \)

\( \beta_j < 0 \Rightarrow P(\text{success}) \downarrow \text{ as } x_j \uparrow \text{ and } P(\text{success}) \uparrow \text{ as } x_j \downarrow \).
9.4.3 Inference

\( \beta_0 \) and \( \beta \) are estimated by Maximum Likelihood:

\[
L(\beta_0, \beta) = \prod_{i=1}^{n} \theta_i^{y_i} (1 - \theta_i)^{1-y_i} ;
\]

\[
\ln L(\beta_0, \beta) = \sum y_i \ln \{\theta_i/(1-\theta_i)\} + \sum \ln (1-\theta_i)
\]

\[
\ln L(\beta_0, \beta) = \ell(\beta_0, \beta) = \sum y_i (\beta_0 + \beta' x_i) - \sum \ln [1+ \exp(\beta_0 + \beta' x_i)]
\]

Standard iterative methods (e.g. Newton-Raphson) give m.l.e.'s \( \hat{\beta}_0, \hat{\beta} \)

\[
\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \theta_i) ; \quad \frac{\partial \ell}{\partial \beta_i} = \sum_{i=1}^{n} x_i (y_i - \theta_i)
\]

Estimated standard errors of these estimates can be obtained from the diagonal of the estimated variance matrix

\[
\text{var}\left(\hat{\beta}_0, \hat{\beta}\right) \approx \left\{-E\left[\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta} \right]\right\}^{-1}_{\hat{\beta}_0, \hat{\beta}}
\]
R or MINITAB or SAS or SPSS or S-PLUS will fit the model and give estimates and standard errors. We can test significance in terms of:

a) **partial z-test**

$$H_0: \beta_j = 0$$

$$\frac{\hat{\beta}_j}{\sqrt{\text{var}(\hat{\beta}_j)}}$$ with N(0,1) %-points

(usually ignore strict need for t-test)

b) **likelihood ratio**

compare $$2 \times |\ell_{\text{full model}} - \ell_{\text{reduced model with } \beta=0}|$$ with $$\chi^2_1$$

where $$\ell$$ is the maximized log likelihood (or deviance)
9.4.4 Example (Pocock p.219)

A trial to assess the effect of the treatment clofibrate on ischaemic heart disease (IHD). Subjects were men with high cholesterol, randomized into placebo and treatment groups.

Prognostic factors (i.e. factors which also affect risk of IHD and which can be identified in advance) were:

- age;
- smoking;
- father’s ‘history’;
- systolic BP;
- cholesterol

Response: \( Y_i \): ‘success’ (!!) = patient subsequently suffers IHD

Each patient has a certain probability \( p_i \) of achieving a response. \( p_i \) is the probability of getting IHD. Define the following multiple logistic model for how \( p_i \) depends on the prognostic variables:

\[
\ln \left( \frac{p_i}{1 - p_i} \right) = \ln \left( \frac{P[\text{suffers IHD}]}{P[\text{does not}]} \right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_6 x_{i6}
\]

where \( \beta_0, \ldots, \beta_6 \) are numerical constants called logistic coefficients.

This is sometimes written logit\( (p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_6 x_{i6} \).

\( x_{i1} = 0 \) (placebo), 1 (clofibrate)
\( x_{i2} = \ln(\text{age}) \)
\( x_{i3} = 0 \) (non-smoker), 1 (smoker)
\( x_{i4} = 0 \) (father alive), 1 (dead)
\( x_{i5} = \text{systolic BP in mm Hg} \)
\( x_{i6} = \text{cholesterol in mg/dl} \)
Apply maximum likelihood to estimate values of $\beta_i$ (i=0,1,...6):

<table>
<thead>
<tr>
<th>factor</th>
<th>$x_j$</th>
<th>$\beta_j$</th>
<th>z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:treatment</td>
<td>0=placebo, 1=treatment</td>
<td>-0.32</td>
<td>-2.9</td>
</tr>
<tr>
<td>2:age</td>
<td>ln(age)</td>
<td>3.0</td>
<td>6.3</td>
</tr>
<tr>
<td>3:smoking</td>
<td>0=non-smok, 1=smoker</td>
<td>0.83</td>
<td>6.8</td>
</tr>
<tr>
<td>4:father’s hist</td>
<td>0=alive, 1=dead</td>
<td>0.64</td>
<td>3.6</td>
</tr>
<tr>
<td>5:systolic BP</td>
<td>Systolic BP in mm Hg</td>
<td>0.011</td>
<td>3.7</td>
</tr>
<tr>
<td>6:cholesterol</td>
<td>Cholesterol in mg/dl</td>
<td>0.0095</td>
<td>5.6</td>
</tr>
</tbody>
</table>

constant term $\beta_0 = -19.60$

\[ \Phi^{-1}(0.005) = z_{0.005} = -2.58 \]
\[ z_{0.025} = -1.96 \]

Treatment: significant, $p < 0.01$; $\beta_1 < 0$;

Probability of IHD is smaller on treatment than on placebo

Prognostic factors: all five significant ($p < 0.01$); all have positive m.l.e.’s, $\therefore$ probability of IHD increases with age, smoking, ‘poorer heredity’, high blood pressure, high cholesterol.
Another useful way of describing the importance of each factor is to look at **odds ratios**. The odds ratio is approximately equal to the relative risk if the probability of the event is small and consequently the term *relative risk* is often [technically mistakenly] used in this context.

e.g. the odds ratio of getting IHD on clofibrate compared with placebo is the ratio of odds:

\[
\frac{P[Y = 1 | x_i = 1]}{P[Y = 0 | x_i = 1]} \cdot \frac{P[Y = 1 | x_i = 0]}{P[Y = 0 | x_i = 0]}
\]

\[= \exp\{\beta_1\}\]

The estimated odds ratio is \(e^{-0.32} = 0.73 < 1\)
i.e. odds of getting IHD are 27% lower on clofibrate after allowing for the other prognostic factors.

The standard error of \(\beta_1\) is 0.11 (= –0.32/–2.9, but actually obtained direct from diagonal of information matrix [not given here]). So approximate 95% confidence limits for \(\beta_1\) are 

\(-0.32 \pm 2 \times 0.11 = –0.10\) and \(-0.54\). Hence \(\exp\{\beta_1\}\) has 95% confidence limits \(e^{-0.1}\) and \(e^{-0.54} = 0.90\) and 0.58 so that 95% confidence limits for the reduction due to clofibrate in odds of getting IHD are 10% and 42%.

Similar calculations for smoking show 95% limits for the increase in odds of getting IHD for smokers are 80% and 193%.
9.4.5 Interactions

Interaction terms would be handled by creating a new variable as the product of the treatment and the covariate values. In the example above the treatment is coded as 0 for placebo and 1 for clofibrate, so the value of this interaction term would be 0 for all subjects receiving placebo and the same as the covariate for those on clofibrate. In the example above Treatment is variable $x_1$ and $\log_e(\text{age})$ is variable $x_2$ and there are six variables in all. We create a new variable $x_7 = x_1 \times x_2$ and then our model is

$$\text{logit}(p_i) = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_6 x_{i6} \text{ for placebo, and}$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + (\beta_2 + \beta_7) x_{i2} + \beta_3 x_{i3} + \ldots + \beta_6 x_{i6} \text{ for clofibrate}$$

and $\beta_7$ reflects the interaction effect, (note that $x_7$ is identical to $x_2$ for those on clofibrate but 0 for those on placebo).

Exactly the same method is appropriate for handling interactions between two continuous covariates and between two 2-level factors. Interactions involving a k-level factor can only be handled by converting the factor into $k-1$ dummy binary variables. In this case the interaction term has $k-1$ degrees of freedom if it is a $k$-level factor $\times$ covariate interaction or $(k-1)(j-1)$ degrees of freedom for an interaction between a $k$-level and a $j$-level factor. This also means that the separate parts of the chi-squared statistic must be combined before assessing significance.
9.4.6 Combining Trials

Within the context of combining trials we might keep $\beta_1$ the same in each trial, but allow $\beta_0$ to vary to reflect possible differences in trial $j$ conditions:

\[
\ln \left( \frac{P[Y_{ij} = 1]}{P[Y_{ij} = 0]} \right) = \beta_j + \beta_1 x_{ij}
\]

i.e.

\[
\ln \left( \frac{P[Y_{ij} = 1]}{P[Y_{ij} = 0]} \right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}
\]

e.g. 3 clinics

\[
\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}
\]

where the last two terms are the clinic coding $x_{i2}$ and $x_{i3}$ are *dummy variables*, i.e.

\[
(x_{i2}, x_{i3}) = (0,0) \text{ for clinic 1} \\
(1,0) \text{ for clinic 2} \\
(0,1) \text{ for clinic 3}
\]

which gives

$\beta_0 + \beta_1 x_{i1}$ for clinic 1

$(\beta_0 + \beta_2) + \beta_1 x_{i1}$ for clinic 2

$(\beta_0 + \beta_3) + \beta_1 x_{i1}$ for clinic 3
9.5 Summary and Conclusions

- Care needs to be taken in analysing matched pairs binary responses. McNemar’s test uses only the information from **unlike pairs**

- Logistic Regression allows the **log-odds** to be modelled as a linear model in the covariates.

- Logistic models can be implemented in most standard statistical packages

- Logistic models allow **relative risks** to be estimated (including confidence intervals).

- Positive coefficients in a logistic model indicate that the factor increases the risk of the ‘success’
Exercises 4

1) Several studies have considered the relationship between elevated blood glucose levels and occurrence of heart problems. The results of two similar studies are summarized below.

<table>
<thead>
<tr>
<th>Glucose Level</th>
<th>Heart Problems</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevated</td>
<td>Yes</td>
<td>61</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1284</td>
<td>996</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>1345</td>
<td>1028</td>
</tr>
<tr>
<td>Not Elevated</td>
<td>Yes</td>
<td>82</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1930</td>
<td>633</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>2012</td>
<td>658</td>
</tr>
<tr>
<td></td>
<td></td>
<td>143</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3214</td>
<td>1629</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3357</td>
<td>1686</td>
</tr>
</tbody>
</table>

i) What can be concluded from these data regarding the influence of glucose on heart problems?

ii) Do you have any doubts on the validity of the form of analysis you have used?
2) A randomized, parallel group, placebo controlled trial was undertaken to assess the effect on children of a cream in reducing the pain associated with venepuncture at the induction of anaesthesia. A binary response of \( Y=0 \) for ‘did not hurt’ and \( Y=1 \) for ‘hurt’ was recorded for each of the 40 children who entered the trial, together with the treatment given \((x_1)\) and two covariates, sex \((x_2)\) and age \((x_3)\), which were thought might affect pain levels. A logistic model was fitted and the following details are available.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Reg. Coeff.</th>
<th>Standard Error of Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.058</td>
<td>1.917</td>
</tr>
<tr>
<td>( x_1 ): treatment ((0 = \text{placebo}, 1 = \text{cream}))</td>
<td>-1.543</td>
<td>0.665</td>
</tr>
<tr>
<td>( x_2 ): sex ((0 = \text{boy}, 1 = \text{girl}))</td>
<td>0.609</td>
<td>0.872</td>
</tr>
<tr>
<td>( x_3 ): age (years)</td>
<td>-0.461</td>
<td>0.214</td>
</tr>
</tbody>
</table>

i) Interpret and assess the treatment effect and also the effects of sex and age.

ii) Estimate the relative risk of hurting with the cream compared to the placebo.
10. Comparing Methods of Measurement

10.1 Introduction

Many situations arise where two (or more) techniques have been used to measure some quantity on the same subject. For example, a new instrument for measuring blood pressure is introduced and compared with an old instrument by taking simultaneous measurements on the same subjects. Another example is when two (or more) observers rate some feature by assigning a category (e.g. good/medium/bad). The first requires the comparison of methods on the basis of continuous measurements, the second on the basis of categorical methods. It would be inappropriate (i.e. wrong) to base the analyses on calculating a correlation coefficient or a $\chi^2$-statistic for independence. In the first case you expect there to be a strong correlation between the measurements on the two instruments and it is of no interest at all whether the correlation is ‘significantly different from zero’. In the second, you already know that the categorizations cannot be independent so it is of no interest to calculate a test of independence. Of much more interest is whether there is some consistent bias by one instrument with respect to the other (does it consistently provide a higher reading?) or whether the observers shew reasonable agreement or not. The two techniques used in these contexts are ‘Bland & Altman Plots’ and calculation of the ‘Kappa statistic’. Neither of these produce any statistical assessment and it is a \textit{clinical decision} whether the degree of agreement is acceptable or not, not a statistical one.

An invaluable reference for this topic are is Martin Bland’s webpage at

http://www-users.york.ac.uk/~mb55/.
10.1 Bland & Altman Plots

The table below, using data from Bland (2000) available from the website referenced above, gives the PEFR in litres/min of 17 subjects measured by two instruments, a Wright meter and a Mini meter.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Wright</th>
<th>Mini</th>
<th>Mean</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>490</td>
<td>525</td>
<td>507.5</td>
<td>-35</td>
</tr>
<tr>
<td>2</td>
<td>397</td>
<td>415</td>
<td>406.0</td>
<td>-18</td>
</tr>
<tr>
<td>3</td>
<td>512</td>
<td>508</td>
<td>510.0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>401</td>
<td>444</td>
<td>422.5</td>
<td>-43</td>
</tr>
<tr>
<td>5</td>
<td>470</td>
<td>500</td>
<td>485.0</td>
<td>-30</td>
</tr>
<tr>
<td>6</td>
<td>611</td>
<td>625</td>
<td>618.0</td>
<td>-14</td>
</tr>
<tr>
<td>7</td>
<td>415</td>
<td>460</td>
<td>437.5</td>
<td>-45</td>
</tr>
<tr>
<td>8</td>
<td>431</td>
<td>390</td>
<td>410.5</td>
<td>41</td>
</tr>
<tr>
<td>9</td>
<td>638</td>
<td>642</td>
<td>640.0</td>
<td>-4</td>
</tr>
<tr>
<td>10</td>
<td>429</td>
<td>432</td>
<td>430.5</td>
<td>-3</td>
</tr>
<tr>
<td>11</td>
<td>420</td>
<td>420</td>
<td>420.0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>633</td>
<td>605</td>
<td>619.0</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>275</td>
<td>227</td>
<td>251.0</td>
<td>48</td>
</tr>
<tr>
<td>14</td>
<td>492</td>
<td>467</td>
<td>479.5</td>
<td>25</td>
</tr>
<tr>
<td>15</td>
<td>165</td>
<td>268</td>
<td>216.5</td>
<td>-103</td>
</tr>
<tr>
<td>16</td>
<td>372</td>
<td>370</td>
<td>371.0</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>421</td>
<td>443</td>
<td>432.0</td>
<td>-22</td>
</tr>
</tbody>
</table>

Comparison of two methods of measuring PEFR (from Bland, 2000)

The next figure is a scatterplot of the two measurements. The line is not the regression line (this would not be appropriate) but the line of equality, i.e. the ideal line if the two instruments agreed perfectly with each other.
There is a suggestion that there are more points above the line than below it but this is not easy to see. More effective is a Bland & Altman Plot which plots the difference against the average of the two measurements. The mean of the differences is \(-9.9\) with standard deviation 36.54, so a 95% confidence interval for the mean difference is \((-27.6, 7.8)\). The difference is a measure of the bias between the two measuring methods so there could be a bias of as much as 28.7 litres per minute. Whether this is unacceptably large is a clinically question, not a statistical issue. Also shown on the graph are what are conventionally known as the limits of agreement which is the mean difference \(\pm 2\times\)standard deviation of differences, i.e. \(-9.9 \pm 2 \times 36.54\) and can be thought of as an approximate 95% confidence interval for an individual difference between the measurements made by the instrument. (The narrower interval calculated above is a 95% confidence interval for the mean difference, i.e. over a long run of measurements.)
Note that Bland & Altman plots do not shew which instrument is the more accurate (they may both be wrong!) but only whether they agree between themselves. It is possible that one of the methods is ‘The Gold Standard’ and the other is a cheaper or more convenient alternative. It is then up to the clinicians involved to decide whether the alternative is acceptably close to the gold standard.
10.2 The Kappa Statistic for Categorical Variables

Suppose two observers rate objects into a set of categories. The kappa statistic is based upon comparing the observed proportion of agreement \( A_{\text{obs}} \) between the two observers with the proportion of agreement \( A_{\text{exp}} \) expected purely by chance. The kappa statistic is then defined as

\[
\kappa = \frac{A_{\text{obs}} - A_{\text{exp}}}{1 - A_{\text{exp}}}
\]

This statistic is *not* assessed in statistical terms but there is a conventional scale of interpretation:

- \( \kappa > 0.75 \):—— excellent agreement
- \( 0.4 < \kappa < 0.75 \):—— fair to good agreement
- \( \kappa < 0.4 \):—— moderate or poor agreement.

The observed agreements are those down the diagonal of the two-way table of assessments made by the two observers and so the observed proportion of agreements is the total of the diagonals divided by the overall total. The expected numbers of agreements are the expected diagonal terms calculated as the product of the marginal totals divided by the overall total (as done in calculating the expected numbers for a chi-squared test on a contingency table).
10.3 Examples

10.3.1 Two Categories

The table below gives the classifications of 179 people who were classified on two occasions as normalizers or non-normalizers after completing a Symptom Interpretation Questionnaire (source: Kirkwood & Stone, 2003).

<table>
<thead>
<tr>
<th>First classification</th>
<th>Normalizer</th>
<th>Non-normalizer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalizer</td>
<td>76</td>
<td>17</td>
<td>93</td>
</tr>
<tr>
<td>Non-normalizer</td>
<td>39</td>
<td>47</td>
<td>86</td>
</tr>
<tr>
<td>Total</td>
<td>115</td>
<td>64</td>
<td>179</td>
</tr>
</tbody>
</table>

The 'expected' agreements are given by (where, e.g. $30.7 = \frac{86 \times 64}{179}$)

<table>
<thead>
<tr>
<th>First classification</th>
<th>Normalizer</th>
<th>Non-normalizer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalizer</td>
<td>59.7</td>
<td></td>
<td>93</td>
</tr>
<tr>
<td>Non-normalizer</td>
<td></td>
<td>30.7</td>
<td>86</td>
</tr>
<tr>
<td>Total</td>
<td>115</td>
<td>64</td>
<td>179</td>
</tr>
</tbody>
</table>

So $A_{obs} = \frac{76+47}{179} = 0.687$ and $A_{exp} = \frac{59.7+30.7}{179} = 0.505$ and so $\kappa = \frac{0.687 - 0.505}{1 - 0.505} = 0.37$ indicating perhaps only moderate agreement.
10.3.2 More than Two Categories

For several categories essentially the same method applies. The table below (Kirkwood & Stone, 2003) give the classification of dominant style of 179 people on two occasions.

<table>
<thead>
<tr>
<th></th>
<th>Normalizer</th>
<th>Somatizer</th>
<th>Psycholgizer</th>
<th>None</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalizer</td>
<td>76</td>
<td>0</td>
<td>7</td>
<td>10</td>
<td>93</td>
</tr>
<tr>
<td>Somatizer</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Psycholgizer</td>
<td>17</td>
<td>1</td>
<td>15</td>
<td>8</td>
<td>41</td>
</tr>
<tr>
<td>None</td>
<td>20</td>
<td>3</td>
<td>5</td>
<td>11</td>
<td>39</td>
</tr>
<tr>
<td>Total</td>
<td>115</td>
<td>4</td>
<td>30</td>
<td>30</td>
<td>179</td>
</tr>
</tbody>
</table>

Calculating \( A_{\text{obs}} = \frac{76+0+15+11}{179} \) = 0.57

The ‘expected’ numbers of interest are:

<table>
<thead>
<tr>
<th></th>
<th>Normalizer</th>
<th>Somatizer</th>
<th>Psycholgizer</th>
<th>None</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalizer</td>
<td>59.7</td>
<td></td>
<td></td>
<td></td>
<td>93</td>
</tr>
<tr>
<td>Somatizer</td>
<td></td>
<td>0.1</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Psycholgizer</td>
<td></td>
<td></td>
<td>6.9</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td></td>
<td>6.5</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>Total</td>
<td>115</td>
<td>4</td>
<td>30</td>
<td>30</td>
<td>179</td>
</tr>
</tbody>
</table>

Giving \( A_{\text{exp}} = 0.409 \) and \( \kappa = 0.27 \) indicating poor agreement.

Note that as the number of categories increases the value of \( \kappa \) is likely to decrease since there are more ‘opportunities’ for misclassification.
10.4 Further Modifications

Two modifications to the kappa statistic are possible but which are not detailed here. The first is when there are several ordered categories where it may be felt that there is a partial agreement for cases classified as only one or two categories apart rather than several. In this case the proportion of agreement could be modified by allowing such partial agreements to contribute to the total with less weight. This could be useful for comparative purposes with other $\kappa$ values calculated with the same system of weighting but does not provide any absolute measure of agreement.

The second modification is when there are more than two observers. In this case an average of all the pairwise $\kappa$-values will provide an overall measure of consistency within the group of observers but there are other possibilities.
10.5 Summary and Conclusions

- It is not appropriate to calculate a correlation coefficient between two methods of measurement to assess the degree of agreement or reproducibility.

- It is appropriate to plot the difference in measurements against their average. This is termed a Bland & Altman plot.

- Levels of agreement are given by
  \[ \text{mean difference} \pm 2 \times \text{st.dev(differences)} \]

- It is not appropriate to calculate a chi-squared statistic for a two-way table of results from two observers to assess the level of agreement.

- A kappa statistic measures the level of agreement.
  
  - 0 < \kappa < 0.4 \Rightarrow \text{poor to moderate agreement},
  
  - 0.4 < \kappa < 0.75 \Rightarrow \text{fair to good agreement},
  
  - 0.75 < \kappa \Rightarrow \text{excellent agreement}.

- Extensions to ordered categories and several observers are possible.
Notes & Solutions for Tasks 1

1) Read the article referred to in §1.8, this can be accessed from the web address given there or from the link given in the course web pages. Use the facility on the BMJ web pages to find related articles both earlier and later.

Trust you have done this by now.

2) Revision of t-tests and non-parametric tests. And this also.

3) Using your general knowledge compare the following two theories against the Bradford-Hill Criteria:

i) Smoking causes lung cancer

Most of the criteria are satisfied. The weakest is whether or not there is a confounding factor that predisposes someone to smoke and that also increases the likelihood of developing lung cancer, possibly genetic. Establishing this criterion can be difficult in the absence of randomised controlled trials (out of the question with humans). The arguments against in this case are that there is evidence of passive smoking being harmful, clear evidence of links between smoking and other diseases (both other forms of cancer and non-cancer conditions such as heart disease), evidence of a link between chewing tobacco and cancers in site topically affected by tobacco juice (mouth and throat in particular).

ii) The MMR (mumps, measles and rubella) vaccine given to young babies causes autism in later childhood.
This theory falls on several criteria. Firstly in terms of consistency, extensive studies in other countries have failed to find evidence of such a connection. In particular a very extensive study in Finland (I leave you to trace an account of this, try googlescholar and also Ben Goldacre’s Bad Science web page). Secondly, specificity is not easy to establish, thirdly no plausible biological mechanism explanation has been offered.
Notes & Solutions for Tasks 2

1) For each of the proposed trials listed below, select the most appropriate study design, allocating onne design to onne trial. (Onne≡’one and only one’!)

A→b
B→a
C→d
D→c

is the best allocation subject to the constraint of onne design used onnce. Some other design might be appropriate for the situation described, e.g. C→a.

2) In a recent radio programme an experiment was proposed to investigate whether common garden snails have a homing instinct and return to their ‘home territory’ if they are moved to some distance away.. The proposal is that you should collect a number of snails, mark them with a distinctly coloured nail varnish, and place all of them in your neighbour’s garden. Your neighbour should do likewise (using a different colour) and place their snails in your garden. You and your neighbour should each observe how many snails returned to their own garden and how many stayed in their neighbour’s. Full details are given at http://downloads.bbc.co.uk/radio4/so-you-want-to-be-a-scientist/Snail-Swapping-Experiment-Instructions.pdf

(a) What flaws does the design of this experiment have?

(b) How could the design of the experiment be improved?

(Note: this question is open-ended and there are many possible acceptable answers to both parts. Discussion is intended)
This question was set in the context of the discussion in lectures of randomized double-blind controlled trials. So the first steps are to consider what the experimental and control groups and what is the ‘intervention’ (i.e. the action performed by the experimenter on the test subjects which might affect the measured outcome — the intervention is performed on the experimental group but not on the control group). In this case the intervention is to move snails from their home territory and place them at some distance. The measured response is to see whether they return to their home territory. Examination of the design shows that there is no control group. This is a major flaw in the design of the experiment. All of the snails caught in the owner’s home garden are marked and placed in the neighbour’s garden. Further, all of the snails marked by the neighbour in there garden are removed to the owner’s garden. If the neighbour marked their snails and then released them back in their own garden then this would be a control group (since they would not have received the intervention). Without this control group you cannot rule out with any certainty whether snails always wander around quite a large territory covering adjacent gardens (remember the time scale is quite long – a week – between intervention and measurement of response).

A further, maybe less serious flaw, is that there is little randomization in the experiment. Presumably the snails that
were captured and marked were not randomly selected from all of those in the garden but were those that were out and about and not hiding in obscure places. It is not realistic to catch all the snails in the garden and select a random sample to be exiled next door. However, a better design would be to catch say 2N snails in the owner’s garden, randomly select N of them to be marked with one colour and then exiled next door, the other N would be marked with a different colour and allowed to stay at home. The neighbour could reciprocate with 2M snails, using two further colours. This would allow control of further potential explanatory factors such as whether snails naturally drift in one direction along the road or whether one garden is particularly attractive to snails because of the presence of young green plants in only one of the gardens and these giving off aromatic signals detectable by snails. If snails equally migrate home in both directions and none of the control groups migrate then it does suggest that the homing instinct is because of homesickness rather than seeking food or some other attraction.
A further design issue is the question of blinding. It would be too easy to bias the results at the point of measurement of response towards a desired outcome by ['subconsciously' or otherwise] not collecting snails marked with the ‘wrong’ colour. Better would be for an independent third party who does not know the colour coding to collect all the marked snails they can find.

The results are given on
http://www.bbc.co.uk/radio4/features/so-you-want-to-be-a-scientist/experiments/homing-snails/results/

Results

<table>
<thead>
<tr>
<th></th>
<th>Totnes Garden Returned to:</th>
<th>Cornwall Campus Returned to:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collected from:</td>
<td>14H, 26A, 10m</td>
<td>54H, 74A, 30m</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fisher's Exact Test: p = 0.0001</td>
<td>Fisher's Exact Test: p = 0.0014</td>
<td></td>
</tr>
<tr>
<td>Collected from:</td>
<td>10H, 68A, 8m</td>
<td>81H, 68A, 8m</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fisher's Exact Test: p = 0.0014</td>
<td>Fisher's Exact Test: p = 0.000001</td>
<td></td>
</tr>
</tbody>
</table>

Key: H,A = number of home and away snails; m = distance between bases; Fisher’s test gives probability of getting our results by chance alone. Small p-values confirm homing instinct.

Findings in this experiment were complicated by a spell of exceptionally dry weather, during which many snails disappeared - presumably in shade and sealed up in their epiphragms. But in those instances where snails were recovered over short distances (up to 10 metres), there was again strong evidence of homing instinct. Over longer distances, particularly over 30 metres, results were inconclusive. This could have been due to the many variables: terrain, e.g. a wood; the type of barrier: e.g. road, building; the hot weather; or the actual distance itself.
This suggests the analysis presented was a Fisher’s exact test (an alternative to a $\chi^2$ test of independence) of a $2\times2$ contingency table, ignoring the fact that few of the snails marked were later found (especially in the ‘Cornwall Campus’). A better analysis is invited from you.

3) **On a recent BBC Radio programme** (Front Row, Friday 03/10/08, [http://www.bbc.co.uk/radio4/arts/frontrow/](http://www.bbc.co.uk/radio4/arts/frontrow/)) there was an interview with Bettany Hughes, a historian, ([http://www.bettanyhughes.co.uk/](http://www.bettanyhughes.co.uk/)) who was talking about gold (in relation to an exhibition of a gold statue of Kate Moss in the British Museum). She made the surprising statement

"...ingesting gold can cure some forms of cancer."

I would only regard this as true if there has been a randomized controlled clinical trial where one of the treatments was gold taken by mouth and where the measured outcome was cure of a type of cancer. The task is to find a record of such a clinical trial or else find a plausible source that might explain this historian’s rash statement.

The basis of this story seems to be reports that gold nano particles have been observed to bind to receptors on certain types of cancer cells. This is a long way from saying that gold *cures cancer*. Looking on clinicaltrials.gov and searching under ‘gold’ ‘cancer’ lists 80+ trials which include the two words ‘gold’ and ‘cancer’ somewhere in their protocols. Several of these use ‘gold’ in the phrase ‘gold standard’ and don’t involve administering actual gold. Others seem to involve studies where gold is not claimed to be the active agent but used as a delivery vehicle for some therapeutic agent bound to colloidal gold (gold pulverised to a very fine powder). I wasn’t able to find details of a couple of Phase I trials (e.g. by Mayo Clinic) but no later phases and no links to publications were given.
4) **What evidence is there that taking fish oil helps schoolchildren concentrate?**

In summary the answer is very little evidence if any at all. A quick search on Ben Goldacre’s page should lead you quickly to this article [http://www.badscience.net/2010/06/the-return-of-a-2bn-fishy-friend/#more-1675](http://www.badscience.net/2010/06/the-return-of-a-2bn-fishy-friend/#more-1675) which tells much of the story. In short, this theory has been reported widely in many newspapers (including recently The Observer, a generally well-regarded Sunday Newspaper) as proven fact. Tracing the Observer article to its source reveals that the study referred to did not involve fish oil nor was it designed to test whether it helped schoolchildren concentrate. It is salutary reading.
Notes & Solutions for Tasks 3

1) Patients are to be allocated randomly to 3 treatments. Construct a randomization list

i) for a simple, unrestricted random allocation of 24 patients

ii) for a restricted allocation stratified on the following factors with 4 patients available in each factor combination:

Sex: M or F

Age: <30; 30≤&<50; ≥50.

i) e.g. take 1,2,3 → A; 4,5,6 → B; 7,8,9 → C; 0 → discard. Or in R:

```r
> x <- c("A", "B", "C")
> y <- sample(x, 24, replace=TRUE)
> y
[1] "C" "B" "B" "A" "A" "A" "A" "A" "A" "A" "C" "B" "C" "C" "C"
"A" "A" "C" "B" "B" "A"
[20] "B" "C" "B" "A" "A"
```

iii) Would usually take 1→ABC; 2→ACB; 3→BAC; 4→BCA; 5→CAB; 6→CBA using randomly permuted blocks of size 3. However, there are only 4 patients available at each factor combination. Possibilities are to choose 4th treatment (a) randomly or (b) selecting if one treatment is more important than the other 2 — then position that treatment randomly in the sequence (4 possible positions). Other possibilities are available.

More sophisticated in R is either:

```r
lapply(rep(list(LETTERS[1:3]),4),sample)
[[1]]
[1] "B" "C" "A"

[[2]]
[1] "B" "A" "C"

[[3]]
[1] "A" "B" "C"

[[4]]
[1] "B" "C" "A"
```

or
2) Patients are to be randomly assigned to active and placebo treatments in the ratio 2:1. To ensure ‘balance’ a block size of 6 is to be used. Construct a randomisation list for a total sample size of 24.

There 15 (=6!/4!2!) blocks of size six of form AAAAPP. Note that a block size of 3 gives only 3 possibilities and so is unsatisfactory – too easy to crack. This can be done easily in R with `rep()` and `sample()`:

```r
> sample(c(rep("A",4),rep("P",2)),6)
> sample(c(rep("A",4),rep("P",2)),6)
> sample(c(rep("A",4),rep("P",2)),6)
[1] "P" "A" "A" "A" "A" "P"
> sample(c(rep("A",4),rep("P",2)),6)
```

More sophisticated is

```r
matrix(apply(matrix(c(rep("A",4),rep("P",2)),6,4),2,sample),1,6*4)
[1,] "A" "A" "A" "P" "A" "P" "A" "A" "A" "A" "P" "P"
[,13] [,,14]
[1,] "A" "A" "A" "P" "A" "A" "A" "P" "A" "P" "A" "P"
[1,] "A" "A" "A" "P" "A" "A" "A" "P" "A" "A" "A" "P"
```

3) Patients are to be randomly assigned to active and placebo treatments in the ratio 3:2. To ensure ‘balance’ a block size of 5 is to be used. Construct a randomisation list for a total sample size of 30.

There are 10 (=5!/3!2!) blocks of size 5 of form AAAPP. Note that a block size of 10 of form AAAPPAAAPP would give 10!/6!4! = 210 possibilities, perhaps too many (overkill), 10 possibilities with block size 5 is probably adequate and not easy to crack, or else take random subset of these of say 5 sets.

Either use repeatedly:

```r
c(sample(c(rep("A", 3), rep("P", 2)), 5)

[1] "A" "A" "P" "A" "P"
```

Or, more sophisticated

```r
matrix(apply(matrix(c(rep("A", 3), rep("P", 2)), 5, 6), 2, sample), 1, 5*6)

[1,] "A" "A" "A" "P" "P" "A" "A" "P" "A" "A" "P" "A" "P"


[1,] "P" "A" "P" "A"
```

4) i) Fifteen individuals who attend a weightwatchers’ clinic are each to be assigned at random to one of the treatments A, B, C to reduce their weights. Describe and implement a randomized scheme to make a balanced allocation of treatments to individuals.
If using a printed table of random numbers (e.g. Neave, Table 7.1) then number people 01, . . . , 15. Take 2-digit random numbers, discard those not between 01 and 15 (fold, to make selection more efficient, if you want; then 01=21=41=61=81, etc); ignore repeats; the first 5 picked get A. Take 5 further 2-digit random numbers between 01 and 15 in the same way; ignore repeats and those that have A; these get B. The remaining 5 get C.

Taking the following random digits (Neave 7.1, row 20):
07636 04876 61063 57571 69434 14965 20911 73162
Take in pairs, fold, so 01=21=41=61=81, etc. 07, 63=03, 60=20 (ignore), 48=08, 76=16 (ignore), 61=01, 06. So: 07, 03, 08, 01, 06 get A. 35=15, 75=15 (ignore), 71=11, 69=09, 43=03 (ignore), 41=01 (ignore), 49=09 (ignore) 65=05, 20 (ignore), 91=11 (ignore), 17 (ignore) 31=11 (ignore) 62=02. So 15, 11, 09, 05, 02 get B. The rest get C.

If using a computer package that has a random number generator or random sample selection then there are various methods. Two are illustrated in R:

(a) \[ x<-c(1:15) \]

\[ y<-sample(x) \]
Then subjects 8, 6, 1, 14 and 5 are allocated to A.

(b)

> z<-c(rep("A",5),rep("B",5),rep("C",5))
> z
[1] "A" "A" "A" "A" "B" "B" "B" "C" "C" "C" "C" "C"
> w<-sample(z)
> w
[1] "B" "C" "A" "A" "C" "A" "B" "B" "A" "C" "B" "C"

Then the first subject is allocated to B, the second to C, etc.

ii) Different individuals need to lose differing amounts of weight—as shown below (in pounds).

<table>
<thead>
<tr>
<th></th>
<th>1. 27</th>
<th>4. 33</th>
<th>7. 27</th>
<th>10. 24</th>
<th>13. 35</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>35</td>
<td>5. 23</td>
<td>8. 34</td>
<td>11. 30</td>
<td>14. 36</td>
</tr>
<tr>
<td>3.</td>
<td>24</td>
<td>6. 26</td>
<td>9. 30</td>
<td>12. 39</td>
<td>15. 30</td>
</tr>
</tbody>
</table>

Describe and implement a design which makes use of this extra information, and explain why this may give a more illuminating comparison of the treatments.

Need to form blocks of similar units (here individuals); ideally, block size is the number of treatments to be compared, so here three. Hence, construct five blocks of size three. Order individuals by weight loss, and then form groups of three, giving the following blocks of individuals: (5, 3, 10), (6, 1, 7), (9, 11, 15), (4, 8, 2), (13, 14, 12); note that '2' and
‘13’ could be the other way round. Now assign each treatment once within each block randomly. Assign an integer to each possible order of the three treatments: 1–ABC, 2–ACB, 3–BAC, 4–BCA, 5–CAB, 6–CBA.

Taking the following random digits (Neave 7.1, row 20): 07636 04876 61063; ignoring 0, 7, 8, 9 gives 6, 3, 6, 4, 6, and so the treatments are assigned in the order: CBA BAC CBA BCA CBA. Comparisons within blocks are made over more similar individuals, thereby reducing the effect on the spread of the results of the external variable ‘how much weight you need to lose’.

In R this could be achieved in a variety of ways, either with allowing different blocks to have the same order of treatments or (since only five of the six possible orderings are required) ensuring that any order is used at most once. Four are illustrated below.

```r
> x<-c(1:6)
> sample(x,5)
[1] 4 1 5 6 3
> sample(x,5,replace=TRUE)
[1] 1 3 4 3 3
> y<- c("ABC", "ACB", "BAC", "BCA", "CAB", "CBA")
> sample(y,5)
```

© NRJF, 1996 → 219
5) A surgeon wishes to compare two possible surgical techniques for curing a specific heart defect, the current standard and a new experimental technique. 24 patients on the waiting list have agreed to take part in the trial; some information about them is given in the table below.

<table>
<thead>
<tr>
<th>Patient</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Age</td>
<td>64</td>
<td>65</td>
<td>46</td>
<td>70</td>
<td>68</td>
<td>52</td>
<td>54</td>
<td>52</td>
<td>75</td>
<td>55</td>
<td>50</td>
<td>38</td>
</tr>
<tr>
<td>Patient</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>Sex</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>F</td>
<td>M</td>
</tr>
<tr>
<td>Age</td>
<td>59</td>
<td>56</td>
<td>64</td>
<td>64</td>
<td>41</td>
<td>68</td>
<td>48</td>
<td>63</td>
<td>41</td>
<td>62</td>
<td>49</td>
<td>44</td>
</tr>
</tbody>
</table>

Devise a suitable way of allocating patients to the two treatments, and carry out the allocation.

There are lots of possible designs; randomization is vital, and balance is important (and easy to obtain). To take advantage of the extra information given, pair the patients up (because there are two treatments) as far as possible by sex and age—since both factors could affect the suitability of the treatment. The female pairs correspond to ages 38 and 46, 49 and 50, 56 and 64, 64 and 65, 68 and 70, or patient numbers 12 and 3, 23 and 11, etc. Similar pairings should be carried out for the males. Within each pair, randomize the two treatments. For example, look up digits from the beginning of Neaves table of random digits: if a pair gets a digit that is odd, assign the standard treatment to the first
patient and the experimental one to the other; if they get an even digit, assign treatments the other way round.

To do this in R we need six randomly selected pairs of AB or BA:

```r
> sample(c("AB","BA"),6,replace=T)
[1] "AB" "BA" "AB" "BA" "AB" "BA"
```

>
Notes & Solutions for Tasks 4

(in all cases take the significance level as 0.05)

The commands in R for calculation of power, sample size etc are `power.t.test()` and `power.prop.test()`. Note that typing the ↑ recalls the last R command and use of Backspace and the ← key allows you to edit the command and run a new version.

1) A trial for the relief of pain in patients with osteoarthritis of the knee is being planned on the basis of a pilot survey which gave a 25% placebo response rate against a 45% active treatment response rate.

i) How many patients will be needed to be recruited to a trial which in a two-sided test will detect a difference of this order of magnitude with 90% power? (Calculate this first ‘by hand’ and then using a computer package and compare the answers).

```r
> power.prop.test(p1=0.25,p2=0.45,power=0.9,sig.level=0.05)

Two-sample comparison of proportions power calculation

     n = 117.4307
p1 = 0.25
p2 = 0.45
sig.level = 0.05
power = 0.9
alternative = two.sided

NOTE: n is number in *each* group
```

So take 118 in each group.

Note that a significance level of 0.05 is assumed by default.

For comparison, the formula gives 115 patients in each group (230 in total), Both Minitab 13 and the program power.exe give 118 (total 236). S-plus 6 gives the same answer to the problem which ever way you feed in the two proportions, the answer it gives is 128. This is the ‘Yates continuity-corrected’ value which is the default option in S-
plus; changing this default in the options panel also gives 118 per group.

**ii)** With equal numbers in placebo and active groups, what active rates would be detected with power in the range 50% to 95% and group sizes 60 to 140? (Calculate for power in steps of 15% and group sizes in steps of 20).

The program power.exe gives the following table

Results
-------

Two Sample test for proportions

Table of CRD calculations

<table>
<thead>
<tr>
<th>Sample size group 1</th>
<th>60</th>
<th>80</th>
<th>100</th>
<th>120</th>
<th>140</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.41887 : 0.39489 : 0.37872 : 0.36689 : 0.35777 :</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>0.45375 : 0.42488 : 0.40536 : 0.39106 : 0.38003 :</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>0.49491 : 0.46048 : 0.43708 : 0.41990 : 0.40661 :</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>0.56566 : 0.52249 : 0.49275 : 0.47073 : 0.45362 :</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rows are: power
significance level = 0.05
ratio group1:group2 = 1:1
group1 proportion = .25

Note the obvious feature that the CRD decreases towards the top-right corner (large sample sizes, low power). This would be used to see what the chances were of detecting a range of differences for some realistic sample size and the benefits in moving to a larger sample size (at perhaps extra cost).

To do this in R without 20 separate calls to `power.prop.test` requires a little bit of programming but can be done quite easily.

```r
> group<-seq(60,140,by=20)
> power<-seq(0.50,0.95,by=0.15)
> group
[1]  60  80 100 120 140
> power
```
[1] 0.50 0.65 0.80 0.95
> delta<-matrix(nrow=4,ncol=5)
> for (i in 1:4) {
+ for (j in 1:5) {
+ delta[i,j]<-power.prop.test(p1=0.25,power=power[i],
+ n=group[j])$p2
+ }
+ }
> options(digits=3)
> delta
[1,] 0.419 0.395 0.379 0.367 0.358
[2,] 0.454 0.425 0.405 0.391 0.380
[3,] 0.495 0.460 0.437 0.420 0.407
[4,] 0.566 0.522 0.493 0.471 0.454
>
2) Woollard & Cooper (1983) Clinical Trials Journal, 20, 89-97, report a clinical trial comparing Moducren and Propranolol as initial therapies in essential hypertension. These authors propose to compare the change in initial blood pressure under the two drugs.

i) Given that they can recruit only 100 patients in total to the study, calculate the approximate power of the two-sided 5% level t-test which will detect a difference in mean values of 0.5σ, where σ is the common standard deviation.
> power.t.test(n=50,sd=1,delta=.5)

Two-sample t test power calculation

n = 50
delta = 0.5
sd = 1
sig.level = 0.05
power = 0.6968888
alternative = two.sided

NOTE: n is number in *each* group
Note that the sample size in each group is 50 (total 100). Also note that a CRD of ½σ means you enter the standard deviation as 1.0 and the CRD as ½.

The programme power.exe gives a value for the power of 69.69%. (The formula for the approximation may give a slightly different answer).
ii) How big a sample would be needed in each group if they required a power of 95%? (Calculate this first ‘by hand’ and then using a computer package and compare the answers).

> power.t.test(power=0.95, sd=1, delta=.5)

```
Two-sample t test power calculation

  n = 104.9280
  delta = 0.5
  sd = 1
  sig.level = 0.05
  power = 0.95
  alternative = two.sided
NOTE: n is number in *each* group
```

Programme power.exe gives 105 in each group (210 in total).

3) Look at the solutions to Task sheet 3 and repeat the analyses given there (if you have not already done so).

   Trust you have done this by now

4) How many subjects are needed to achieve a power of 80% when the standard deviation is 1.5 to detect a difference in two populations means of 0.8 using a two sample t-test? (Note that R gives the number needed in each group, i.e. total is twice number given)

> power.t.test(sd=1.5, power=.8, delta=0.8)

```
Two-sample t test power calculation

  n = 56.16413
  delta = 0.8
  sd = 1.5
  sig.level = 0.05
  power = 0.8
  alternative = two.sided
NOTE: n is number in *each* group
```

So we need 57 in each group (note we need to round fractional sample sizes up to nearest integer) and therefore 114 in total.
5) How many subjects are needed to achieve a power of 80% when the standard deviation is 1.5 to detect a difference in one population mean from a specified value of 0.8 using a one sample t-test?

```r
> power.t.test(sd=1.5,power=.8,delta=0.8,type="one.sample")
```

One-sample t test power calculation

<table>
<thead>
<tr>
<th>n</th>
<th>delta</th>
<th>sd</th>
<th>sig.level</th>
<th>power</th>
<th>alternative</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.57195</td>
<td>0.8</td>
<td>1.5</td>
<td>0.05</td>
<td>0.8</td>
<td>two.sided</td>
</tr>
</tbody>
</table>

Thus we need 30 subjects.

6) Do you have an explanation for why the total numbers in Q2 and Q3 are so different?

Some people might think that if you need N for specified power and delta with a one sample test then you need 2N for a two sample test but in fact you will need about 4N. My personal 'explanation/visualisation' of what is happening is that with two samples each sample mean can be either above or below the target population mean – it is only when they are both as far away from the other population mean as possible that the strongest evidence of a difference in population means is provided. This is only one of the four possible combinations of whether the two sample means are above or below their population means. Perhaps a more technical explanation is that two variances have to be estimated rather than only one.

7) How many subjects are needed to detect a change of 20% from a standard incidence rate of 50% using a two sample test of proportions with a power of 90%?

```r
> power.prop.test(power=.9,p1=.5,p2=.7)
```

Two-sample comparison of proportions power calculation

<table>
<thead>
<tr>
<th>n</th>
<th>p1</th>
</tr>
</thead>
<tbody>
<tr>
<td>123.9986</td>
<td>0.5</td>
</tr>
</tbody>
</table>
p2 = 0.7
sig.level = 0.05
power = 0.9
alternative = two.sided

NOTE: n is number in *each* group

> power.prop.test(power=.9,p1=.5,p2=.3)

Two-sample comparison of proportions power calculation

n = 123.9986
p1 = 0.5
p2 = 0.3
sig.level = 0.05
power = 0.9
alternative = two.sided

NOTE: n is number in *each* group
Note that it does not matter whether the change from .5 is up or down. Rounding up we see we need 124 in each group so 248 in total.

8) How many subjects are need to detect a change from 30% to 10% using a two sample test of proportions with a power of 90%?

power.prop.test(power=.9,p1=.1,p2=.3)

Two-sample comparison of proportions power calculation

n = 81.96206
p1 = 0.1
p2 = 0.3
sig.level = 0.05
power = 0.9
alternative = two.sided

NOTE: n is number in *each* group
So we need 164 in total.

9) How many subjects are needed to detect a change from 60% to 80% using a two sample test of proportions with a power of 90%?

> power.prop.test(power=0.9,p1=.6,p2=.8)

Two-sample comparison of proportions power calculation

n = 108.2355

So we need 218 in total
10) How many subjects are needed to detect a change from 50% to 30% using a two sample test of proportions with a power of 90%?

You should have answered this in Q5

11) How many subjects are needed to detect a change from 75% to 55% using a two sample test of proportions with a power of 90%?

> power.prop.test(power=0.9,p1=.75,p2=.55)

Two-sample comparison of proportions power calculation  n = 117.4307

So 236 in total.

12) How many subjects are needed to detect a change from 40% to 60% using a two sample test of proportions with a power of 90%?

> power.prop.test(power=0.9,p1=.4,p2=.6)

Two-sample comparison of proportions power calculation  n = 129.2529

So 260 in total.

13) Questions 5, 6, 7, 8, 9 and 10 all involve changes of 20% and a power of 90%. Why are the answers not all identical?

It is because when estimating a proportion as the number of success r out of n trials the standard error of the estimate is \( \sqrt{\frac{r}{n}(1-\frac{r}{n})/n} \) which is a maximum when \( \frac{r}{n}=\frac{1}{2} \), i.e. proportions closer to 0.5 require a greater sample size for a specified precision than those further from 0.5.

14) Without doing any calculations (neither by hand nor in R) write down the number of subjects needed to detect a change from 45% to 25% using a two sample test of proportions with a power of 90%.

236 in total (same as Q11).
Notes & Solutions for Tasks 5

1) Senn and Auclair (Statistics in Medicine, 1990, 9) report on the results of a clinical trial to compare the effects of single inhaled doses of 200μg salbutamol (a well established bronchodilator) and 12μg formoterol (a more recently developed bronchodilator) for children with moderate or severe asthma. A two-treatment, two-period crossover design was used with 13 children entering the trial, and the observations of the peak expiratory flow, a measure of lung function where large values are associated with good responses, were taken. The following summary of the data is provided.

<table>
<thead>
<tr>
<th>Group 1: formoterol → salbutamol ( (n_1 = 7) )</th>
<th>Group 2: salbutamol → formoterol ( (n_2 = 6) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period 1</td>
<td>Period 2</td>
</tr>
<tr>
<td>mean</td>
<td></td>
</tr>
<tr>
<td>337.1</td>
<td>306.4</td>
</tr>
<tr>
<td>s.d.</td>
<td></td>
</tr>
<tr>
<td>53.8</td>
<td>64.7</td>
</tr>
</tbody>
</table>

a) Specify a model for peak expiratory flow which incorporates treatment, period and carryover effects.

Model: usual one in notes. It is a good idea to plot the means for each group for each period (not shewn here) and then see that it is suggestive that treatment 2 is superior, no obvious carryover nor period effects.
b) Assess the carryover effect, and, if appropriate, investigate treatment differences. In each case specify the hypotheses of interest and illustrate the appropriateness of the test.

Carryover: \( t = 0.17 \ \text{[} (643.6 - 629.2) / (114.3^2/7 + 174^2/6)^{1/2} \text{]} \) \( p > 0.05 \), so can proceed with treatment & period tests:

Treatment: \( t = 4.22 \ \text{[} (30.7 - (-62.6)) / (33.0^2/7 + 44.7^2/6)^{1/2} \text{]} \) on 6 d.f., \( p < 0.01 \), so clear evidence of a difference between the treatments.

Inspection of the means shews that formoterol is superior.

Period: \( t = -1.44 \) (on 6 df), \( p = 0.2 \), no evidence of a systematic difference between periods.

(demonstrate appropriateness of tests by reference to model as in notes).

Conclude that there is strong evidence that formoterol gives a better response than salbutamol.
2) A and B are two hypnosis treatments given to insomniacs one week apart. The order of receiving the treatment is randomized between patients. The measured response is the number of hours sleep during the night. Data are given in the following table.

<table>
<thead>
<tr>
<th>patient</th>
<th>period 1</th>
<th>period 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>B</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
<td>3</td>
</tr>
</tbody>
</table>

b) Calculate the mean for each treatment in each period and display the results graphically.

b) Assess the carryover effect.

c) If appropriate, assess the treatment and period effects.

(NB These data are available in R, Minitab and S-PLUS forms on the course web pages)

Given below is a transcript of R performing all the required calculations using the command `t.test(.)`.
The relevant values and key steps needed to answer the questions above have been highlighted in the transcript below. Note the slick trick used to change the signs of the group 2 differences. This is not something you actually need to be able to do yourself, just recognise it later.

> hourssleep

<table>
<thead>
<tr>
<th>PERIOD1</th>
<th>PERIOD2</th>
<th>GROUP</th>
<th>sum</th>
<th>diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>14</td>
<td>2</td>
<td>12.5</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>8</td>
<td>2</td>
<td>10.0</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>8.0</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>6.0</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>4.0</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3.5</td>
</tr>
<tr>
<td>9</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>7.5</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>7.0</td>
</tr>
<tr>
<td>13</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>4.5</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>5.5</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>7.0</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>5.0</td>
</tr>
</tbody>
</table>

> attach(hourssleep)

> t.test(sum[GROUP==1],sum[GROUP==2])

Welch Two Sample t-test

data:  sum[GROUP == 1] and sum[GROUP == 2]
t = -0.9929, df = 12.64, p-value = 0.3394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-4.176408  1.51408
sample estimates:
mean of x  mean of y
5.3750   6.6875
> t.test(diff[,GROUP==1],diff[,GROUP==2])
  Welch Two Sample t-test
  data:  diff[,GROUP == 1] and diff[,GROUP == 2]
  t = 2.3503, df = 11.543, p-value = 0.03746
  alternative hypothesis: true difference in means is not
equal to 0
  95 percent confidence interval: 
  0.3703012 10.3796988
  sample estimates:
  mean of x mean of y
  5.500     0.125

SLICK TRICK HERE<<<<<<<<<<<<<<<<<<<<<<<<<<<<!!!!!!!

> treatindicator<-3-2*unclass(GROUP)
> treatindicator
  [1]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
  attr(,"levels")
  [1] "1" "2"
> treatdiff<-diff*treatindicator
> treatdiff
  [1]  9  3 -4 -4  0 10  0  1 11 -4 -1 12  3  1  2  4
  attr(,"levels")
  [1] "1" "2"
> t.test(treatdiff[,GROUP==1],treatdiff[,GROUP==2])
  Welch Two Sample t-test
  data:  treatdiff[,GROUP == 1] and treatdiff[,GROUP == 2]
  t = 2.4597, df = 11.543, p-value = 0.03077
  alternative hypothesis: true difference in means is not
equal to 0
  95 percent confidence interval: 
  0.6203012 10.6296988
  sample estimates:
  mean of x mean of y
  5.500     -0.125

<table>
<thead>
<tr>
<th></th>
<th>Group 1: A→ B (n₁ = 8)</th>
<th>Group 2: B→ A(n₂ = 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Period 1</td>
<td>Period 2</td>
</tr>
<tr>
<td>mean</td>
<td>8.13</td>
<td>2.625</td>
</tr>
<tr>
<td>s.d.</td>
<td>4.29</td>
<td>2.50</td>
</tr>
</tbody>
</table>
The following R code will produce a 'nice' plot of mean responses but it is probably sufficient in most routine cases to produce a quick one by hand.

```r
> GP1PER1mean <- mean(PERIOD1[GROUP==1])
> GP1PER2mean <- mean(PERIOD2[GROUP==1])
> GP2PER1mean <- mean(PERIOD1[GROUP==2])
> GP2PER2mean <- mean(PERIOD2[GROUP==2])
> per <- c(1,2)
> gp1 <- c(GP1PER1mean, GP1PER2mean)
> gp2 <- c(GP2PER1mean, GP2PER2mean)
> ymax <- max(GP1PER1mean, GP1PER2mean, GP2PER1mean, GP2PER2mean)
> ymin <- min(GP1PER1mean, GP1PER2mean, GP2PER1mean, GP2PER2mean)
> ymax <- ymax + 0.1 * (ymax - ymin)
> ymin <- ymin - 0.1 * (ymax - ymin)
> plot(xlim=c(0.9,2.1), ylim=c(ymin, ymax), type="n", xlab="period",
+ ylab="mean hours sleep", xaxt="n",
+ main="Plot of mean responses against periods")
> axis(1, at=c(1,2))
> points(per, gp1, pch=15, col="blue", cex=1.5)
> points(per, gp2, pch=16, col="red", cex=1.5)
> lines(per, gp1, col="blue", lwd=2)
> lines(per, gp2, col="red", lwd=2)
> gp1labels <- c("Treat A", "Treat B")
> text(per, gp1, labels=gp1labels, adj=c(.9,1.4))
> gp2labels <- c("Treat B", "Treat A")
> text(per, gp2, labels=gp2labels, adj=c(.9,1.4))
```

![Plot of mean responses against periods](image-url)
Note that plot suggests that A is better than B and that there is a period effect (the average results in period 2 are lower than those in period 1). Whether there is a carryover effect is a more difficult matter of judgement. If there is carryover then it is quite complex and not only is B persisting to depress the results on A for group 2 but A is interacting with B to produce substantially lower results in period 2 for group 1. It would be surprising that such and interaction would be so different for the two groups. A simpler explanation (i.e. use Occam’s Razor) is that it is a combination of period and treatment effects. This is not contradicted by the formal statistical tests. These are (taking values from output — though you could do this from the summary statistics in the table above using the two sample t-test used in the first question, though with a conservative d.f. = 8 rather than R’s calculated values of 11 or 12).

Carryover: $t = -0.99$, d.f. = 12, $p = 0.340$, no evidence.

Period: $t = 2.46$, d.f. = 11, $p = 0.032$, good evidence of difference in periods.

Treatment: $t = 2.35$, d.f. = 11, $p = 0.038$, good evidence that A is better than B.
Notes & Solutions for Tasks 6

1) Two ointments A and B have been widely used for the treatment of athlete's foot. In a recent report the following results were noted, where response indicated temporary relief from the outbreak.

<table>
<thead>
<tr>
<th>Response</th>
<th>No Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ointment A</td>
<td>174</td>
</tr>
<tr>
<td>Ointment B</td>
<td>149</td>
</tr>
</tbody>
</table>

a) Based on these results the report concluded that ointment A was more effective than ointment B. Use the Mantel-Haenszel test to verify this conclusion.

b) Further investigation into the source of the data revealed that the data had been pooled from two clinics. The results from individual clinics were:

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Ointment A</th>
<th>Ointment B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Response</td>
<td>No response</td>
</tr>
<tr>
<td>1</td>
<td>129</td>
<td>71</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>25</td>
</tr>
</tbody>
</table>

Reassess the evidence in the light of these additional facts.

Use the formulae in §8.3.

Overall: \(E[Y_1]=161.5\), \(\text{var}(Y_1)=32.50\), \(\chi^2_{\text{MH}}=4.8\); \(p<0.05\)

Clinic 1: \(E[Y_1]=121.0\), \(\text{var}(Y_1)=23.96\), \(\chi^2_{\text{MH}}=2.67\); \(p>0.05\)

Clinic 2: \(E[Y_1]=40.5\), \(\text{var}(Y_1)=8.59\), \(\chi^2_{\text{MH}}=2.36\); \(p>0.05\)

Conclude that there is very strong evidence that A is more effective. (response rates are 64.5%, and 64.3% — very close, so few doubts on validity of combining results.)
Below is a complete analysis in R:

```r
> x<factor(rep(c(1,2),c(200,200)),labels=c("Oint A","Oint B"))
> y<factor(rep(c(1,2,1,2),c(129,71,113,87)),labels=c("Response","No Response"))
> z<factor(rep(1,400),labels="Clinic 1")
> table(x,y,z)

<table>
<thead>
<tr>
<th>x</th>
<th>Response</th>
<th>No Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oint A</td>
<td>129</td>
<td>71</td>
</tr>
<tr>
<td>Oint B</td>
<td>113</td>
<td>87</td>
</tr>
</tbody>
</table>

> mantelhaen.test(x,y,z,correct=F)

Mantel-Haenszel chi-squared test without continuity correction

data:  x and y and z
Mantel-Haenszel X-squared = 2.6714, df = 1, p-value = 0.1022
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
0.9353062 2.0921389
sample estimates:
common odds ratio
1.398853

> x<factor(rep(c(1,2),c(70,70)),labels=c("Oint A","Oint B"))
> y<factor(rep(c(1,2,1,2),c(45,25,36,34)),labels=c("Response","No Response"))
> z<factor(rep(1,140),labels="Clinic 2")
> table(x,y,z)

<table>
<thead>
<tr>
<th>x</th>
<th>Response</th>
<th>No Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oint A</td>
<td>45</td>
<td>25</td>
</tr>
<tr>
<td>Oint B</td>
<td>36</td>
<td>34</td>
</tr>
</tbody>
</table>

> mantelhaen.test(x,y,z,correct=F)

Mantel-Haenszel chi-squared test without continuity correction

data:  x and y and z
Mantel-Haenszel X-squared = 2.3559, df = 1, p-value = 0.1248
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
0.8635901 3.3464951
sample estimates:
common odds ratio
1.7

> x<factor(rep(c(1,2,1,2),c(200,200,70,70)),
```

© NRJF, 1996 → 237
Solutions to Tasks.

+ labels=c("Oint A","Oint B"))
> y<-factor(rep(c(1,2,1,2,1,2,1,2),
+ c(129,71,113,87,45,25,36,34)),
+ labels=c("Response","No Response"))
> z<-factor(rep(c(1,2),c(400,140)),
+ labels=c("Clinic 1","Clinic 2"))
> table(x,y,z)

, , z = Clinic 1

Y
 x          Response No Response
Oint A     129          71
Oint B     113          87

, , z = Clinic 2

Y
 x          Response No Response
Oint A      45          25
Oint B      36          34

> mantelhaen.test(x,y,z,correct=F)

Mantel-Haenszel chi-squared test without continuity correction
data:  x and y and z
Mantel-Haenszel X-squared = 4.7999, df = 1, p-value = 0.02846
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
1.041550 2.080194
sample estimates:
common odds ratio
1.471946

2) (Artificial data from Ben Goldacre, 06/08/11).

Imagine a study was conducted to examine the relationship between heavy drinking of alcohol and developing ling cancer, obtaining the following results:

<table>
<thead>
<tr>
<th>Cancer</th>
<th>No Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drinker</td>
<td>366</td>
</tr>
<tr>
<td>Non-Drinker</td>
<td>98</td>
</tr>
</tbody>
</table>

c) Calculate the ratio of the odds of developing cancer for drinkers to non-drinkers. What conclusions do you draw from this odds ratio?
The odds ratio is 3.01, suggesting that the odds for developing cancer are three times higher for drinkers than for non-drinkers. An approximate 95% confidence interval for the odds ratio is (2.38, 3.81).

d) It transpires that 330 of the drinkers developing cancer were smokers and 1100 of the drinkers who smoked did not, with corresponding figures for the non-drinkers of 47 and 156. Calculate the odds ratios separately for smokers and non-smokers. What conclusions do you draw?

Both the odds ratios are 1.0, suggesting that the key difference in cancer rates is between smokers and non-smokers with no evidence of a difference between drinkers and non-drinkers. This effect is essentially the same as that observed in Simpson’s paradox and illustrates the danger of post-hoc regrouping of tables. See the original article at http://www.guardian.co.uk/commentisfree/2011/aug/05/bad-science-adjusting-figures
Notes & Solutions for Exercises 1

1) In the comparison of a new drug A with a standard drug B it is required that patients are assigned to drugs A and B in the proportions 3:1 respectively. Illustrate how this may be achieved for a group of 32 patients, and provide an appropriate randomization list. Comment on the rationale for selecting a greater proportion of patients for drug A.

(i) Need blocks of form AAAB (or of form AAAAAABB). There are 4 of form AAAB (and 28 of size 8). Using 1,2→AAAB; 3,4→AABA; 5,6→ABAA; 7,8→BAAA, 9,0→ignore, a sequence of random digits 7,1,4,2,0,1,8,1,2,4 gives BAAA|AAAB|AABA|AAAB|AAAB|BAAA|AAAB|AAAB.

In R, to produce a random block of form AAAB do:

```r
> sample(c(rep("A",3),"B"))
[1] "A" "A" "B" "A"
```

and then repeat as often as necessary or build into a loop.

Alternatively, to get exact balance without blocks do:

```r
> sample(c(rep("A",24),rep("B",8)))
[29] "A" "B" "A" "A"
```

There could be economic reasons for using more As than Bs, but more likely if B is the standard then there will be interest in efficacy and safety of the new treatment but this is likely to be known for the standard, as would be drop out rates, standard deviations etc. Having more patients on the new treatment protects against uncertainty in drop-out rates (or side effects) and consistency of response. Further, there will be more interest and enthusiasm amongst both patients and investigators if there is a greater chance of receiving the new treatment and so easier to recruit centres and patients. This last
reason is probably the most important in practice though not obviously ‘statistical’.

2) The table below gives the age (≤55/>55), gender (M/F), disease stage (I/II/III) of subjects entering a randomized controlled clinical trial at various intervals and who are to be allocated to treatment or placebo in approximately equal proportions immediately on entry.

<table>
<thead>
<tr>
<th>order of entry</th>
<th>Age</th>
<th>Gender</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>≤55</td>
<td>F</td>
<td>III</td>
</tr>
<tr>
<td>2</td>
<td>≤55</td>
<td>M</td>
<td>III</td>
</tr>
<tr>
<td>3</td>
<td>≤55</td>
<td>M</td>
<td>I</td>
</tr>
<tr>
<td>4</td>
<td>≤55</td>
<td>F</td>
<td>I</td>
</tr>
<tr>
<td>5</td>
<td>&gt;55</td>
<td>F</td>
<td>II</td>
</tr>
<tr>
<td>6</td>
<td>≤55</td>
<td>F</td>
<td>III</td>
</tr>
<tr>
<td>7</td>
<td>&gt;55</td>
<td>F</td>
<td>I</td>
</tr>
<tr>
<td>8</td>
<td>&gt;55</td>
<td>M</td>
<td>III</td>
</tr>
<tr>
<td>9</td>
<td>≤55</td>
<td>M</td>
<td>III</td>
</tr>
<tr>
<td>10</td>
<td>&gt;55</td>
<td>F</td>
<td>III</td>
</tr>
<tr>
<td>11</td>
<td>≤55</td>
<td>F</td>
<td>III</td>
</tr>
<tr>
<td>12</td>
<td>≤55</td>
<td>M</td>
<td>I</td>
</tr>
<tr>
<td>13</td>
<td>&gt;55</td>
<td>F</td>
<td>I</td>
</tr>
</tbody>
</table>

i) Construct a randomization list for this group of subjects by a minimization method designed to achieve an overall balance between the factors.

<table>
<thead>
<tr>
<th>order of entry</th>
<th>Age</th>
<th>Gender</th>
<th>Stage</th>
<th>First Run</th>
<th>Second Run</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>score for T</td>
<td>score for P</td>
</tr>
<tr>
<td>1</td>
<td>≤55</td>
<td>F</td>
<td>III</td>
<td>0**</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>≤55</td>
<td>M</td>
<td>III</td>
<td>2</td>
<td>0**</td>
</tr>
<tr>
<td>3</td>
<td>≤55</td>
<td>M</td>
<td>I</td>
<td>1**</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>≤55</td>
<td>F</td>
<td>I</td>
<td>4</td>
<td>1**</td>
</tr>
<tr>
<td>5</td>
<td>&gt;55</td>
<td>F</td>
<td>II</td>
<td>1</td>
<td>1**</td>
</tr>
<tr>
<td>6</td>
<td>≤55</td>
<td>F</td>
<td>III</td>
<td>4**</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>&gt;55</td>
<td>F</td>
<td>I</td>
<td>3**</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>&gt;55</td>
<td>M</td>
<td>III</td>
<td>4</td>
<td>3**</td>
</tr>
<tr>
<td>9</td>
<td>≤55</td>
<td>M</td>
<td>III</td>
<td>6**</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>&gt;55</td>
<td>F</td>
<td>III</td>
<td>7</td>
<td>6**</td>
</tr>
<tr>
<td>11</td>
<td>≤55</td>
<td>F</td>
<td>III</td>
<td>9</td>
<td>8**</td>
</tr>
<tr>
<td>12</td>
<td>≤55</td>
<td>M</td>
<td>I</td>
<td>8</td>
<td>6**</td>
</tr>
</tbody>
</table>
The first subject has to be allocated randomly to T or P. The ★ indicates which of T or P is selected. Then for each subsequent subject it is easy to calculate the score for T and P as the total number of characteristics held in common between the new arrival and those subjects already allocated to that group. Two runs are presented above, one resulting from a choice of T for the first subject — this leads to a tied score for the 5th subject and P was [randomly] chosen, another tie for the 9th and T was [randomly] chosen. The second run with P selected first also leads to a tie on the 5th arrival and then the 9th.

ii) Cross-tabulate the treatment received with each [separate] factor.

Run 1:

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Gender</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≤55</td>
<td>&gt;55</td>
<td>total</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>P</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>total</td>
<td>8</td>
<td>5</td>
<td>13</td>
</tr>
</tbody>
</table>

Run 2:

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Gender</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≤55</td>
<td>&gt;55</td>
<td>total</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>P</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>total</td>
<td>8</td>
<td>5</td>
<td>13</td>
</tr>
</tbody>
</table>
Note that these are identical, as are essentially all possible runs (i.e. up to an interchange of T and P). Even with a different order of arrival of these patients the final allocations are not substantially different.
iii) Construct a list to allocate the subjects to treatment completely randomly without taking any account of any prognostic factor and compare the balance between treatment groups achieved on each of the factors.

In R the function `sample(.)` with the `replace=TRUE` option gives the same facility:

```R
> sample(c("T","P"),13,replace=TRUE)
[1] "T" "P" "T" "T" "T" "T" "P" "P" "T" "P" "T" "P" "T"
```

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤55</td>
<td>&gt;55</td>
<td>total</td>
</tr>
<tr>
<td>T</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>total</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>I</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>III</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>total</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

(Different randomisations will lead to different cross-tabulations.)
Notes & Solutions for Exercises 2

1) In a clinical trial of the use of a drug in twin pregnancies an obstetrician wishes to show a significant prolongation of pregnancy by use of the drug when compared to placebo. She assesses that the standard deviation of pregnancy length is 1.5 weeks, and considers a clinically significant increase in pregnancy length of 1 week to be appropriate.

i) How many pregnancies should be observed to detect such a difference in a test with a 5% significance level and with 80% power?

Require a two-sided two sample t-test. Formula gives 35.3 per group and R, Minitab and programme POWER give 37 in each group (S-PLUS gives 36) so 74 (or 72) pregnancies in total need to be observed.

```
> power.t.test(sd=1.5,delta=1,power=0.8)
Two-sample t test power calculation
 n = 36.3058
delta = 1
 sd = 1.5
 sig.level = 0.05
 power = 0.8
 alternative = two.sided
NOTE: n is number in *each* group
```

ii) It is thought that between 40 and 60 pregnancies will be observed to term during the course of the study. What range of increases in length of pregnancy will the study have a reasonable chance (i.e. between 70% and 90%) of detecting?

Note that “40 to 60 in total” means 20 to 30 in each group.

Results produced by programme POWER below:

Results
-------
Two Sample T test

Table of CRD calculations

<table>
<thead>
<tr>
<th>Sample size group 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>: 20 : 25 : 30 :</td>
</tr>
</tbody>
</table>

| 70 : 1.20670 : 1.07390 : 0.97708 : |
| 75 : 1.27967 : 1.13884 : 1.03617 : |
| 80 : 1.36103 : 1.21125 : 1.10205 : |
| 85 : 1.45595 : 1.29572 : 1.17890 : |
| 90 : 1.57545 : 1.40207 : 1.27566 : |
Rows are: power significance level = 0.05 standard deviation = 1.5
This will give an answer apparently accurate to about 6 seconds (since
the working units are days and so they should be rounded to one (or at
most two) decimal places.

In R, using the routine given in Task Sheet 3 we have

```r
> group <- seq(20, 30, by=5)
> power <- seq(0.70, 0.90, by=0.05)
> group
[1] 20 25 30
> power
[1] 0.70 0.75 0.80 0.85 0.90
> delta <- matrix(nrow=5, ncol=3)
> for (i in 1:5) {
+   for (j in 1:3) {
+     delta[i,j] <- power.t.test(sd=1.5, power=power[i],
+                                n=group[j])$delta
+   }
+ }
> options(digits=3)
> delta
   [,1]   [,2]   [,3]
[1,] 1.21 1.08 0.978
[2,] 1.28 1.14 1.038
[3,] 1.36 1.21 1.103
[4,] 1.46 1.30 1.180
[5,] 1.58 1.40 1.277
```

There are some numerical differences in these but only of the order of
about 10 minutes.
Notes & Solutions for Exercises 3

1) Given below is an edited extract from an SPSS session analysing the results of a two period crossover trial to investigate the effects of two treatments A (standard) and B (new) for cirrhosis of the liver. The figures represent the maximal rate of urea synthesis over a short period and high values are desirable. Patients were randomly allocated to two groups: the 8 subjects in group 1 received treatment A in period 1 and B in period 2. Group 2 (13 subjects) received the treatments in the opposite order.

i) Specify a suitable model for these data which incorporates treatment, period and carryover effects.

ii) Assess the evidence that there is a carryover effect from period 1 to period 2.

iii) Do the data provide evidence that there is a difference in average response between periods 1 and 2?

iv) Assess whether the treatments differ in effect, taking into account the results of your assessments of carryover and period effects.

v) Repeat the statistical analysis in R

vi) The final stage in the analysis recorded below produced 95% Confidence Intervals, firstly, for the mean differences in response between periods 1 and 2 for the 21 subjects and, secondly, for the mean differences in response to treatments A and B for the 21 subjects. By referring to your model for these data, explain why these two confidence intervals can not be used to provide indirect tests of the hypotheses of no period and no treatment effects respectively.

vii) Under what circumstances would the confidence intervals described in part (e) provide valid assessments of period and treatment effects?

A plot of mean responses (not shewn here, but always advisable) indicates that there looks to be a difference between the treatments (with B better) and little suggestion of period or carryover effects. This gives a useful guide to ensuring the t-tests are selected correctly.
i) Usual model from notes, including the identifiability constraints (i.e. sums = 0)
ii) No evidence of carryover (t = .314)
iii) Little evidence of difference in periods (t = 0.49, p = 0.63) (period 1 lower)
iv) Some evidence of treatment differences, t = −2.019, p = 0.059 (using both periods since no evidence of carryover (nor period) effect). mean response to B is higher than to A so some evidence that new treatment is better.

```r
> attach(cirrhosis)
> cirrhosis[1:5,]
Patnum Group Period1 Period2 Sum1.2 PeriodDiffs TreatDiffs
1  1  1     48      51     99     -3      -3
2  2  1     43      47     90     -4      -4
3  3  1     60      66    126     -6      -6
4  4  1     35      40     75     -5      -5
5  5  1     36      39     75     -3      -3
>
> t.test(Sum1.2 ~ Group)

Welch Two Sample t-test

data:  Sum1.2 by Group
t = 0.3137, df = 18.683, p-value = 0.7572
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:  
-15.50916  20.97070
sample estimates:  
mean in group 1 mean in group 2
  93.50000     90.76923
```
Solutions to Exercise

```r
> t.test(PeriodDiffs ~ Group)

Welch Two Sample t-test

data:  PeriodDiffs by Group
t = -2.0192, df = 17.646, p-value = 0.05893
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:   
-12.1340837   0.2494683
sample estimates:   
mean in group 1 mean in group 2    
-2.250000 3.692308

> t.test(TreatDiffs ~ Group)

Welch Two Sample t-test

data:  TreatDiffs by Group
t = 0.4901, df = 17.646, p-value = 0.6301
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:   
-4.749468 7.634084
sample estimates:   
mean in group 1 mean in group 2    
-2.250000 -3.692308

> t.test(PeriodDiffs)

One Sample t-test

data:  PeriodDiffs
t = 0.8863, df = 20, p-value = 0.386
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:   
-1.933623 4.790766
sample estimates:   
mean of x 1.428571
```
> t.test(TreatDiffs)

    One Sample t-test

data:  TreatDiffs
t = -2.116, df = 20, p-value = 0.04709
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  -6.241143  0.04457117
sample estimates:
  mean of x
  -3.142857

v) If you go back to the model and calculate the expected value of
the mean differences (remembering \( \tau_A + \tau_B = 0 \) etc) you find that they
involve both \( \pi_1 - \pi_2 \) and \( \tau_A - \tau_B \) in both cases whereas you would
want to have the expected value to be e.g. just \( \pi_1 - \pi_2 \) for the
Confidence Interval to provide a test of \( \pi_1 - \pi_2 = 0 \) — instead it
provides a CI for that expected value. Specifically, the mean
difference between period 1 and period 2 involves 8 terms of form
\((\mu + \tau_A + \pi_1) - (\mu + \tau_B + \pi_2) = \tau_A - \tau_B + \pi_1 - \pi_2\) and 13 terms of form
\((\mu + \tau_B + \pi_1) - (\mu + \tau_A + \pi_2) = \tau_B - \tau_A + \pi_1 - \pi_2\) (ignoring the \( \alpha \) and \( \epsilon \) terms which have
expectation 0). So the expected mean value of the period
difference is

\[
[8(\tau_A - \tau_B + \pi_1 - \pi_2) + 13(\tau_B - \tau_A + \pi_1 - \pi_2)] / 21 = \pi_1 - \pi_2 - 5(\tau_A - \tau_B) / 21
\]

and so if there is a large treatment effect the CI for this mean
difference could exclude 0 even if there is no period effect. Similar
calculations for the mean treatment difference give parallel
conclusions.

vi) Again, from the calculations you can see that it would be ok if
the sample sizes were equal.
Extract from SPSS Analysis of Crossover Trial on Liver Treatment

Summarize

<table>
<thead>
<tr>
<th>Patnum</th>
<th>Group</th>
<th>Period1</th>
<th>Period2</th>
<th>Sum1+2</th>
<th>Diff1–2</th>
<th>Diff1–(–2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>48.00</td>
<td>51.00</td>
<td>99.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>2.00</td>
<td>1.00</td>
<td>43.00</td>
<td>47.00</td>
<td>90.00</td>
<td>-4.00</td>
<td>-4.00</td>
</tr>
<tr>
<td>3.00</td>
<td>1.00</td>
<td>60.00</td>
<td>66.00</td>
<td>126.00</td>
<td>-6.00</td>
<td>-6.00</td>
</tr>
<tr>
<td>4.00</td>
<td>1.00</td>
<td>35.00</td>
<td>40.00</td>
<td>75.00</td>
<td>-5.00</td>
<td>-5.00</td>
</tr>
<tr>
<td>5.00</td>
<td>1.00</td>
<td>36.00</td>
<td>39.00</td>
<td>75.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>6.00</td>
<td>1.00</td>
<td>43.00</td>
<td>46.00</td>
<td>89.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>7.00</td>
<td>1.00</td>
<td>46.00</td>
<td>52.00</td>
<td>98.00</td>
<td>-6.00</td>
<td>-6.00</td>
</tr>
<tr>
<td>8.00</td>
<td>1.00</td>
<td>54.00</td>
<td>42.00</td>
<td>96.00</td>
<td>12.00</td>
<td>12.00</td>
</tr>
<tr>
<td>9.00</td>
<td>2.00</td>
<td>31.00</td>
<td>34.00</td>
<td>65.00</td>
<td>-3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>10.00</td>
<td>2.00</td>
<td>51.00</td>
<td>40.00</td>
<td>91.00</td>
<td>11.00</td>
<td>-11.00</td>
</tr>
<tr>
<td>11.00</td>
<td>2.00</td>
<td>31.00</td>
<td>34.00</td>
<td>65.00</td>
<td>-3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>12.00</td>
<td>2.00</td>
<td>43.00</td>
<td>36.00</td>
<td>79.00</td>
<td>7.00</td>
<td>-7.00</td>
</tr>
<tr>
<td>13.00</td>
<td>2.00</td>
<td>47.00</td>
<td>38.00</td>
<td>85.00</td>
<td>9.00</td>
<td>-9.00</td>
</tr>
<tr>
<td>14.00</td>
<td>2.00</td>
<td>29.00</td>
<td>32.00</td>
<td>61.00</td>
<td>-3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>15.00</td>
<td>2.00</td>
<td>35.00</td>
<td>44.00</td>
<td>79.00</td>
<td>-9.00</td>
<td>9.00</td>
</tr>
<tr>
<td>16.00</td>
<td>2.00</td>
<td>58.00</td>
<td>50.00</td>
<td>108.00</td>
<td>8.00</td>
<td>-8.00</td>
</tr>
<tr>
<td>17.00</td>
<td>2.00</td>
<td>60.00</td>
<td>60.00</td>
<td>120.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>18.00</td>
<td>2.00</td>
<td>82.00</td>
<td>63.00</td>
<td>145.00</td>
<td>19.00</td>
<td>-19.00</td>
</tr>
<tr>
<td>19.00</td>
<td>2.00</td>
<td>51.00</td>
<td>50.00</td>
<td>101.00</td>
<td>1.00</td>
<td>-1.00</td>
</tr>
<tr>
<td>20.00</td>
<td>2.00</td>
<td>49.00</td>
<td>42.00</td>
<td>91.00</td>
<td>7.00</td>
<td>-7.00</td>
</tr>
<tr>
<td>21.00</td>
<td>2.00</td>
<td>47.00</td>
<td>43.00</td>
<td>90.00</td>
<td>4.00</td>
<td>-4.00</td>
</tr>
</tbody>
</table>

T-Test

Independent Samples Test

<table>
<thead>
<tr>
<th></th>
<th>Mean Difference</th>
<th>Std. Error Difference</th>
<th>t</th>
<th>Df</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum1+2</td>
<td>2.7308</td>
<td>8.7046</td>
<td>.314</td>
<td>18.683</td>
<td>.757</td>
</tr>
<tr>
<td>Diff1–2</td>
<td>-5.9423</td>
<td>2.9429</td>
<td>-2.019</td>
<td>17.646</td>
<td>.059</td>
</tr>
<tr>
<td>Diff1–(–2)</td>
<td>1.4423</td>
<td>2.9429</td>
<td>.490</td>
<td>17.646</td>
<td>.630</td>
</tr>
</tbody>
</table>
### Summarize

#### Case Summaries (a)

<table>
<thead>
<tr>
<th>GROUP</th>
<th>N</th>
<th>Summ1+2</th>
<th>Diff1–2</th>
<th>Diff1–(–2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1</td>
<td>99.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>90.00</td>
<td>-4.00</td>
<td>-4.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>126.00</td>
<td>-6.00</td>
<td>-6.00</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>75.00</td>
<td>-5.00</td>
<td>-5.00</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>75.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>89.00</td>
<td>-3.00</td>
<td>-3.00</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>98.00</td>
<td>-6.00</td>
<td>-6.00</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>96.00</td>
<td>12.00</td>
<td>12.00</td>
</tr>
<tr>
<td>Total</td>
<td>N</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>93.5000</td>
<td>-2.2500</td>
<td>-2.2500</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>16.1688</td>
<td>5.8979</td>
<td>5.8979</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GROUP</th>
<th>N</th>
<th>Diff1–2</th>
<th>Diff1–(–2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.00</td>
<td>1</td>
<td>65.00</td>
<td>-3.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>91.00</td>
<td>11.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>65.00</td>
<td>-3.00</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>79.00</td>
<td>7.00</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>85.00</td>
<td>9.00</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>61.00</td>
<td>-3.00</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>79.00</td>
<td>-9.00</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>108.00</td>
<td>8.00</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>120.00</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>145.00</td>
<td>19.00</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>101.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>91.00</td>
<td>7.00</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>90.00</td>
<td>4.00</td>
</tr>
<tr>
<td>Total</td>
<td>N</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>90.7692</td>
<td>3.6923</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>23.6684</td>
<td>7.4876</td>
</tr>
</tbody>
</table>

### Explore

<table>
<thead>
<tr>
<th></th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff1–2 95% CI for Mean</td>
<td>-1.9336</td>
<td>4.7908</td>
</tr>
<tr>
<td>Diff1–(–2) 95% CI for Mean</td>
<td>-6.2411</td>
<td>-0.044571</td>
</tr>
</tbody>
</table>
Notes & Solutions for Exercises 4

1) Several studies have considered the relationship between elevated blood glucose levels and occurrence of heart problems. The results of two similar studies are summarized below.

<table>
<thead>
<tr>
<th>Glucose Level</th>
<th>Heart Problems</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevated</td>
<td>yes</td>
<td>61</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>1284</td>
<td>996</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1345</td>
<td>1028</td>
</tr>
<tr>
<td>Not Elevated</td>
<td>yes</td>
<td>82</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>1930</td>
<td>633</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2012</td>
<td>658</td>
</tr>
<tr>
<td></td>
<td></td>
<td>143</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3214</td>
<td>1629</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3357</td>
<td>1686</td>
</tr>
</tbody>
</table>

i) What can be concluded from these data regarding the influence of glucose on heart problems?

ii) Do you have any doubts on the validity of the form of analysis you have used?

Mantel-Haenszel tests:

Study 1: $E[Y_1]=1345\times143/3357=57.29$

$\text{var}(Y_1)=1345\times2012\times143\times3214/(3357^2\times3356)=32.89$

so $\chi^2_{\text{MH}}=0.417$, $p>>0.05$.

Study 2: $E[Y_2]=34.75$, $\text{var}(Y_2)=13.11$, $\chi^2_{\text{MH}}=0.579$, $p>>0.05$.

Combined gives $\chi^2_{\text{MH}}=0.02$.

Conclude that there is no evidence of influence of glucose on heart problems. Response rates in the two studies are 4.5% and 3.1%, not very different in absolute terms so few doubts as to validity of analysis, and in any case the results are so far away from significance. Note that the Pearson $\chi^2$ values are nearly identical to the Mantel-Haenszel ones.
Just for illustration, but beyond the scope of this question, here is an analysis using logistic regression: First set up the data as

```r
> frequency<-c(61,82,1284,1930,32,25,996,633)
> problems<-c(rep(c(1,1,0,0),2))
> glucose<-c(rep(c(1,0),4))
> study<-c(rep(0,4),rep(1,4))
>
> heart.glm<-glm(problems~glucose+study,weights=frequency,family=binomial)
>
> summary(heart.glm)
```

**Call:**

```r
glm(formula = problems ~ glucose + study, family = binomial, weights = frequency)
```

**Deviance Residuals:**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-6.558</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Coefficients:**

|          | Estimate | Std. Error | z value | Pr(>|z|) |
|----------|----------|------------|---------|---------|
| (Intercept) | -3.12076 | 0.10426 | -29.933 | <2e-16 *** |
| glucose   | 0.02069  | 0.14737 | 0.140   | 0.888   |
| study     | -0.24457 | 0.16251 | -1.505  | 0.132   |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1682.9 on 7 degrees of freedom
Residual deviance: 1680.6 on 5 degrees of freedom
AIC: 1686.6

Number of Fisher Scoring iterations: 6

>
2) A randomized, parallel group, placebo controlled trial was undertaken to assess the effect on children of a cream in reducing the pain associated with venepuncture at the induction of anaesthesia. A binary response of $Y=0$ for ‘did not hurt’ and $Y=1$ for ‘hurt’ was recorded for each of the 40 children who entered the trial, together with the treatment given ($x_1$) and two covariates, sex ($x_2$) and age ($x_3$), which were thought might affect pain levels. A logistic model was fitted and the following details are available.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Regression Coefficient</th>
<th>Standard Error of Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.058</td>
<td>1.917</td>
</tr>
<tr>
<td>$x_1$: treatment</td>
<td>-1.543</td>
<td>0.665</td>
</tr>
<tr>
<td></td>
<td>(0 = placebo, 1 = cream)</td>
<td></td>
</tr>
<tr>
<td>$x_2$: sex</td>
<td>0.609</td>
<td>0.872</td>
</tr>
<tr>
<td></td>
<td>(0 = boy, 1 = girl)</td>
<td></td>
</tr>
<tr>
<td>$x_3$: age (years)</td>
<td>-0.461</td>
<td>0.214</td>
</tr>
</tbody>
</table>

i) Interpret and assess the treatment effect and also the effects of sex and age.

ii) Estimate the relative risk of hurting with the cream compared to the placebo.

<table>
<thead>
<tr>
<th>Fact or Coefficient</th>
<th>coefficient/s.e.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>-1.543</td>
<td>-2.32</td>
</tr>
<tr>
<td>sex</td>
<td>0.609</td>
<td>0.698</td>
</tr>
<tr>
<td>age</td>
<td>-0.461</td>
<td>-2.15</td>
</tr>
</tbody>
</table>

Good evidence that treatment reduces the relative risk of hurting (or more exactly of children reporting pain). Also good evidence that this risk decreases with age. No evidence of differences between sexes.
Estimate of relative risk using cream is $e^{-1.543} = 0.2137$ or 21.4\%, with an approximate 95\% CI of (5.7\%, 80.8\%). So the reduction in risk when using the cream is estimated as 79\%, with 95\% CI of (19\%, 94\%).